OXFORD

STANDING UP FOR PHILOSOPHY

armchairs, experiments, and the case for methodological reform

Jonathan M. Weinberg & Joshua Alexander

Standing Up for Philosophy

Standing Up for Philosophy

Armchairs, Experiments, and the Case for Methodological Reform

JONATHAN M. WEINBERG
AND
JOSHUA ALEXANDER

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America.

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2025931671

ISBN 9780192856982

DOI: 10.1093/9780191947704.001.0001

Printed and bound by
CPI Group (UK) Ltd., Croydon, CR0 4YY

The manufacturer's authorised representative in the EU for product safety is
Oxford University Press España S.A., Parque Empresarial San Fernando de Henares,
Avenida de Castilla, 2 – 28830 Madrid (www.oup.es/en).

For my Dad, of blessed memory
whose love and faith in me were total and unwavering
(even if he perhaps wished that "experimental philosophy"
involved a few more pipettes and Bunsen burners).
—JW

In loving memory of my mom, who loved books more than anyone I've ever known,
and who would have especially loved this one for reasons obvious to any parent,
and to my dad, who taught me to say "Cogito, ergo sum" at the age of four,
imagining as parents do where that knowledge would take me. It took me here.
—JA

Contents

Acknowledgments

This idea to write this book began on a walk we took in Sedona during a conference in December 2016 that was organized by Ángel Pinillos and hosted by the philosophy department at Arizona State University. But the ideas in this book began over lunch at the Uptown Cafe in Bloomington sometime in the spring of 2002 when we first met and started talking philosophy, a conversation that hasn't stopped for over twenty years. Over the course of the more than twenty years that we have been walking together and talking philosophy, the ideas in this book have been shaped by countless other conversations we have had with too many friends and colleagues to properly thank here. We hope that they will see reflected in this book the many important insights that they have shared with us along the way, and will not see any of the mistakes that they have tried to help us avoid making. Special thanks to some of these friends and colleagues who took the time to read and comment on drafts of this book, including Sara Aronowitz, Sam Bennett, Jessica Brown, Nick Byrd, David Colaço, Nat Hansen, Allan Hazlett, Joshua Knobe, Edouard Machery, Ron Mallon, Aaron Meskin, Jenny Nado, Shaun Nichols, Hannah Rakoski, and Rissa Willis; as well as the students in the University of Arizona methodology seminar in fall 2020. We presented parts of this book at several places during its development, including the joint Hong Kong University/University of Tokyo workshop on Progress in Philosophy in summer 2019, David Rose's Werkmeister X-Phi Conference at Florida State University in fall 2019, a colloquium presentation to Tufts University in fall 2019, and Joachim Horvath's Emmy Noether Research Group on Experimental Philosophy and the Method of Cases in fall 2021. We also presented material at two meetings of the Pacific Division of the American Philosophical Association where we were critics for two authors, Max Deutsch and Edouard Machery, whose books provided much of the inspiration for us to write our own. Parts of Chapters 2 and 3 that address their books were previously printed as stand-alone essays in *Analysis* and *Oxford Studies in Experimental Philosophy*. We would also like to thank the International School of Tucson for the extensive use of its library in the early stages of the preparation of this manuscript. Above all, we are enormously grateful to our families who have always stood up for us and who taught us to stand up for what we believe in.

1
Introduction

This book is a defense of analytic philosophy. It is also a defense of experimental philosophy. In fact, it is a defense of the one in no small part by also being a defense of the other. This might seem odd since the two are often thought to be deeply at odds with one another; so much so, in fact, that it is rather common nowadays to think that experimental philosophers must think that something is terribly wrong with analytic philosophy and that analytic philosophers must think that something is terribly wrong with experimental philosophy. In this spirit, Timothy Williamson (2010) famously wrote in the *New York Times*:

> There are philosophy-hating philosophers who would like to replace the traditional methodology of philosophy, with their stress on a combination of abstract reasoning and particular examples, by something more like imitation psychology. Without even properly defining what it is they are attacking, they use experimental results in a selective and unscientific spirit to try to discredit the traditional methodology.

While Williamson has softened his view of experimental philosophy since then, this way of thinking about the relationship between experimental philosophy and analytic philosophy is still held widely throughout the profession, and our goal is to show, once and for all, that this way of thinking about the relationship between analytic philosophy and experimental philosophy is deeply mistaken.[1] In order to do this, we need to explain the relationship between analytic philosophy and experimental philosophy, including the reasons why experimental philosophy *is* a species of analytic philosophy and the reasons why some philosophers have sometimes thought that it is not. It turns out that what many philosophers think about

[1] To get a better sense for how Williamson now thinks about experimental philosophy, see Williamson (2016). We'll go ahead and lay down a marker right off the bat that we completely agree with his more updated assessment there: "Philosophy cannot be reduced to psychology: no clear or plausible picture of an alternative philosophical method has emerged from experimental philosophers' critique of armchair philosophy. There may indeed be a role for experimental philosophy in refining current philosophical method, but only once the method of experimental philosophy has itself been considerably refined" (p. 35). This book is in part an attempt to meet the challenge set out in these concluding lines of his paper: to articulate such a constructive and progressive role for experimental philosophy in philosophical inquiry more broadly.

Standing Up for Philosophy. Jonathan M. Weinberg and Joshua Alexander, Oxford University Press.
 DOI: 10.1093/9780191947704.003.0001

the relationship between the two depends on the kind of framework that is often used to describe whatever perceived metaphilosophical differences and disagreements there are between them, and so one of our central goals in this book will be to develop and advance a new framework for thinking about them.[2] In order to do this, like any record that needs to be set straight, we need to begin at the beginning and so that is where we will start.

1 The Method of Cases

Is knowledge justified true belief? Are we morally responsible for our actions only if we could have acted otherwise? Are our actions morally permissible just in case they provide the greatest benefit for the greatest number of people all else being equal? Analytic philosophers often try to answer these kinds of questions in significant part by using something called the *method of cases*, a method that involves testing our answers to philosophical questions like these, that is, our philosophical theories themselves, against what we think about real or imagined cases.[3] So, for example, we measure our theories of knowledge against what we think about "Gettier cases" like this one (Gettier 1963, 122):

> Suppose that Smith and Jones have applied for a certain job. And suppose that Smith has strong evidence for the following conjunctive proposition:
>
> (d) Jones is the man who will get the job, and Jones has ten coins in his pocket.
> Smith's evidence for (d) might be that the president of the company assured him that Jones would in the end be selected, and that he, Smith, had counted the coins in Jones's pocket ten minutes ago. Proposition (d) entails:
> (e) The man who will get the job has ten coins in his pocket.

[2] For the sake of linguistic economy we will often use "philosophy" and its cognates when in fact, as indicated here, we very much mean analytic philosophy, and indeed mostly Anglophone analytic philosophy. We do not at all mean thereby to be saying that such philosophy is the only kind of philosophy or anything of that nature. We are simply not in a position to make generalizations to other of the many forms of philosophy worldwide.

[3] It is important to keep in mind that analytic philosophy is a diverse *research community*, to use a term that we will talk about in much more detail in Chapter 4, with a rich set of research interests. Here we are interested in a set of practices paradigmatically associated with what analytic philosophers have sometimes called "philosophical analysis," though it also includes many researchers who do not think of themselves as doing linguistic or conceptual analysis. But even researchers who specialize in the method of cases of course use other methods as well, as we will discuss below when we talk about the resources available to "armchair philosophy," and throughout the book. In fact, one of our main objectives in the book is to convince analytic philosophers that they should adopt a richer set of methodological resources than the ones that are available to them from the "armchair." Also, our discussion of "the" method of cases should not obscure the fact that there are other, non-evidentiary uses to which philosophers often put their discussion of various scenarios; such uses simply fall outside our concern here, and nothing we say should be taken to impugn them (see Maynes 2021).

> Let us suppose that Smith sees the entailment from (d) to (e), and accepts (e) on the grounds of (d), for which he has strong evidence. In this case, Smith is clearly justified in believing that (e) is true.
>
> But imagine, further, that unknown to Smith, he himself, not Jones, will get the job. And, also, unknown to Smith, he himself has ten coins in his pocket. Proposition (e) is then true, though proposition (d), from which Smith inferred (e), is false. In our example, then, all of the following are true: (i) (e) is true, (ii) Smith believes that (e) is true, and (iii) Smith is justified in believing that (e) is true. But it is equally clear that Smith does not *know* that (e) is true; for (e) is true in virtue of the number of coins in Smith's pocket, while Smith does not know how many coins are in Smith's pocket, and bases his belief in (e) on a count of the coins in Jones's pocket, whom he falsely believes to be the man who will get the job.

And we measure our theories of moral responsibility against what we think about "Frankfurt-style cases" like this one (Frankfurt 1969, 835–836):

> Suppose someone—Black, let us say—wants Jones to perform a certain action. Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones is about to make up his mind what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones is going to decide to do something *other* than what he wants him to do. If it does become clear that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do. Whatever Jones's initial preferences and inclinations, then, Black will have his way.
>
> What steps will Black take, if he believes he must take steps, in order to ensure that Jones decides and acts as he wishes? Anyone with a theory concerning what "could have done otherwise" means may answer this question for himself by describing whatever measure he would regard as sufficient to guarantee that, in the relevant sense, Jones cannot do otherwise. Let Black pronounce a terrible threat, and in this way both force Jones to perform the desired action and prevent him from performing a forbidden one. Let Black give Jones a potion, or put him under hypothesis, and in some such way as these generate in Jones an irresistible inner compulsion to perform the act Black wants performed and to avoid others. Or let Black manipulate the minute processes of Jones's brain and nervous system in some more direct way, so that causal forces running in and out of his synapses and along the poor man's nerves determine that he chooses to act and that he does act in the one way and not in the other. Given any conditions under which it will be maintained that Jones cannot do otherwise, in other words, let Black bring it about that those conditions prevail. The structure of the example is

> flexible enough, I think, to find a way around any charge of irrelevance by accommodating the doctrine on which the charge is based.
>
> Now suppose that Black never has to show his hand because Jones, for reasons of his own, decides to perform and does perform the very action Black wants him to perform. In that case, it seems clear, Jones will bear precisely the same moral responsibility for what he does as he would have borne if Black had not been ready to take steps to ensure that he does it. It would be quite unreasonable to excuse Jones for his action, or to withhold the praise to which it would normally entitle him, on the basis of the fact that he could not have done otherwise.

And we measure our theories of moral permissibility against what we think about "trolley cases" like this one (Thomson 1985, 1395–1396):

> Some years ago, Philippa Foot drew attention to an extraordinarily interesting problem. Suppose you are the driver of a trolley. The trolley rounds a bend, and there come into view ahead five track workmen, who have been repairing the track. The track goes through a bit of a valley at that point, and the sides are steep, so you must stop the trolley if you are to avoid running the five men down. You step on the brakes, but alas they don't work. Now you suddenly see a spur of track leading off to the right. You can turn the trolley onto it, and thus save the five men on the straight track ahead. Unfortunately, Mrs. Foot has arranged that there is one track workman on that spur of track. He can no more get off the track in time than the five can, so you will kill him if you turn the trolley onto him. Is it morally permissible for you to turn the trolley?
>
> Everybody to whom I have put this hypothetical case says, Yes, it is. Some people say something stronger than that it is morally *permissible* for you to turn the trolley: They say that morally speaking, you *must* turn it—that morality requires you to do so. Others do not agree that morality requires you to turn the trolley, and even feel a certain discomfort at the idea of turning it. But everybody says that it is true, at a minimum, that you *may* turn it—that it would not be morally wrong for you to do so.

It is beyond question that philosophical practice frequently involves using the method of cases, and deploying as evidence our rendered judgments about these cases. We will call these judgments, when deployed evidentially, *case verdicts* or sometimes simply *verdicts*.[4] So, the verdict that Smith does not know that the person who will get the job has ten coins in his pocket counts as evidence that knowledge is not merely justified true belief, the verdict that Jones is morally

[4] Though it will not be just *any* judgment that can gain this status, as we will argue below. But it's much more important to see in what follows that our focus is ultimately not on the verdicts per se, but on the specific philosophical *practices* that appeal to such verdicts.

responsible for his actions counts as evidence that we can be morally responsible for our actions even when we could not have done otherwise, and so on. What are matters of vigorous, and often heated, philosophical debate are questions about the metaphysical nature and epistemic status of this kind of evidence. What sort of cognition is involved in generating these kinds of case verdicts? Is this practice of appealing to case verdicts a respectable one, like appealing to ordinary perceptual evidence, or is it suspect, like appealing to hunches or tea leaves?

Philosophical Intuitions

Far and away the most common answer to that first question, but one that we want to reject for reasons that we will discuss below, is that the verdicts about philosophical cases are generated by our *philosophical intuitions*, or simply *intuitions*, for short. So, for example, George Bealer (1996, 122) writes about the impact that Gettier's paper had in philosophy:

> Now at one time many people accepted the doctrine that knowledge was justified true belief. But today we have good evidence to the contrary, namely, our intuitions that situations like those described in the Gettier literature are possible and that the relevant people in those situations would not know the things at issue. This and countless other examples show that, according to our standard justificatory procedure, intuitions are used as evidence (or as reasons).

Bealer not only makes the *descriptive* claim that philosophical intuitions are part of our standard justificatory procedure in philosophy, he also argues for the *normative* claim that they must be. This normative claim is part of his famous argument for the incoherence of empiricism, whose basic idea is that empiricists cannot defend their own methodological commitments using any set of justificatory resources that doesn't include our philosophical intuitions. Other philosophers have gone further than Bealer, claiming not only that intuitions are part of our standard justificatory procedure, but also that this fact is part of what makes philosophy methodologically unique. So, for example, Janet Levin (2005, 193–194) writes:

> This procedure of rejecting or modifying theses in the face of intuitively convincing counterexamples has been characteristic, perhaps definitive, of philosophical argumentation throughout its history.

And Alvin Goldman (2007, 2) writes:

> One thing that distinguishes philosophical methodology from the methodology of the sciences is its extensive and avowed reliance on intuition. Especially when

> philosophers are engaged in philosophical "analysis," they often get preoccupied with intuitions. To decide what is knowledge, reference, identity, or causation (or what is the concept of knowledge, reference, identity, or causation), philosophers routinely consider actual and hypothetical examples and ask whether these examples provide instances of the target category or concept. People's mental responses to these examples are often called "intuitions" and these intuitions are treated as evidence for the correct answer. At a minimum, they are evidence for the examples' being instances or non-instances of knowledge, reference, causation, etc. Thus, intuitions play a particularly critical role in a certain sector of philosophical activity.

We could go on and on (and on). The common idea is that the method of cases involves, at least in significant part, advancing philosophical theories on the basis of their ability to explain our intuitions, defending the truth of those theories on the basis of their overall agreement with our intuitions, and justifying our beliefs that those theories are true on the basis of their accordance with our intuitions. Since the method of cases is such a common part of philosophical practice, and since the most standard story about what is supposed to ground this method are our philosophical intuitions about the cases, it is completely natural that much of the recent discussion and debate about philosophical methodology has turned around the metaphysical nature and epistemic status of intuitions. Our aim in this book, nonetheless, is not to contribute to the literature that has built up around questions about the nature and status of philosophical intuitions. Our intention, rather, is to make a sharp left turn to exit from these kinds of discussions. It is not that discussions about the nature and status of intuitions are of no *philosophical* value. Far from it. But the problem, as we hope to make clear in what follows, is that these kinds of discussions have little *methodological* value. This probably shouldn't be surprising. After all, practical methodological questions involving other sources of evidence in other domains almost never turn on debates about metaphysical nature and epistemic status of those sources of evidence, very often because these philosophical debates are over questions that abstract away from the methodological particularities that are most relevant to folks using these sources of evidence in practice. Nothing, for example, about the philosophy of perception is, or was ever meant to be, of any relevance to a radiologist trying to read an x-ray film.

Another reason why it probably shouldn't be surprising that recent debates about the metaphysical nature and epistemic status of intuitions have little methodological value is that these debates have started to feel quite a bit like the latest chapter in the centuries-old, unresolvable debate about rationalism and empiricism, where that debate has often stalled on disagreements about the nature and status of different kinds of evidence, and so have provided little by way of concrete results that could offer positive methodological prescriptions. Consider, for starters, the standing dissensus about what sort of things intuitions even are in the

first place. Even philosophers who all agree that we do and should appeal to intuitions diverge at key points about just what it is that we may be appealing to when we make such appeals. Here's a short list of some of the different sorts of things that different philosophers seem to be talking about when they talk about philosophical intuitions: some philosophers take themselves to be talking about manifestations of our semantic competence, with or without any special phenomenological features (Jackson 1998, Henderson and Horgan 2001; Jackman 2001); some philosophers take themselves to be talking about a form of first-personal access to the contents of our concepts (Goldman 2007); some philosophers take themselves to be talking about deliverances of heuristic cognition (Weatherson 2014); some philosophers take themselves to be talking about products of our capacity for folk-psychological mentalizing (Nagel 2012) or mental modeling, more generally (Alexander 2016); some philosophers take themselves to be talking about rational insights delivered from neo-Aristotelian mental contact with the universals themselves (BonJour 1998); some philosophers take themselves to be talking about empirically derived theories or conceptions (Devitt 2006, Jenkins 2008); some philosophers take themselves to be talking about simply a species of judgment in general (Williamson 2007); and some philosophers take themselves to be talking about the application of refined first-order inferential competencies (Ichikawa and Jarvis 2013). None of this disagreement about the nature of intuitions seems to matter in the slightest to the first-order philosophical debates in which these philosophers are participants.[5]

And we haven't even gotten to the part about all the disagreement there is about what it means to rely on philosophical intuitions as evidence. Does it mean relying on psychological states as evidence, propositions about psychological states as evidence, the contents of psychological states as evidence, or something else? Nor have we gotten to the part about all the disagreement concerning how we should go about trying to figure out what philosophical intuitions are supposed to be. Are we supposed to trust our intuitions about intuitions, pay special attention to what they introspectively seem to be from the first-person point of view, look closely at what philosophers appeal to when they appeal to intuitional evidence in practice, or something else again?

It gets even worse. There is also a wide, and sometimes sharply divergent, set of views about what it is about intuitions that is supposed to make them fit for epistemic service. When they work well, are they deliverances of pure conceptual competence, or judgments usefully informed by background empirical information? Are philosophers doing it right, or doing it wrong, when they let their own philosophical views shape their intuitions—that is, are intuitions supposed to be

[5] We discuss several different methodological challenges that come from adopting different ways of thinking about the metaphysical nature and epistemological status of philosophical intuitions in Alexander and Weinberg (2014).

pristinely pre-theoretic, or to draw upon our substantial philosophical theoretical training? Are they the product of some sort of rationally apperceptive contact with the universals themselves (BonJour 1998) or just a reliable piece of cognitive machinery that has been shaped and tuned by evolution (Nagel 2012, Nagel et al. 2013)? Metaphilosophical questions proliferate here, as well, about just how much naturalistic, psychological evidence should or should not be mustered.

The Method in the Method of Cases

With so little agreement about what they are, what it means to rely on them as evidence, why they might be good candidates for evidence in the first place, and how we are even supposed to go about settling disagreements on these issues, we find ourselves agreeing with Williamson (2007, 220), who suggests that:

> Philosophers might be better off not using the word "intuition" and its cognates. Their main current function is not to answer questions about the nature of the evidence on offer but to fudge them, by appearing to provide answers without really doing so.

And we think that the best way to avoid the thorny issues involved in thinking about intuitional evidence is to focus on the method of cases *as a method* rather than on any particular view about the kinds of mental states that are involved when we think about philosophical cases. To do this, let's return for a moment to Bealer's description of philosophical practice. Like a lot of philosophers whose metaphilosophical focus has been on philosophical intuitions, he runs together two things that can, and should, be treated separately: namely, a philosophical method that involves using our verdicts about real and imagined cases as evidence that certain philosophical theories are true or false, and the kinds of mental states that this method putatively involves. Our target here is the method: the visible, public actions and transactions involved in giving and responding to talks, publishing papers and books, and so on. Setting aside questions about the nature and status of these mental states, what can we say about the method?

Here are some preliminary answers to this question. To be clear, though, before we start, we are *not* offering criteria, let alone necessary and sufficient conditions, for the method of cases. We take the method to be one that can be easily picked out ostensively by analytic philosophers familiar with the sort of work we invoked at the start of this chapter. What we are offering here are some substantive, if rough, generalizations and observations about that method as practiced today.

Thus: for any proposed case *C* and verdict about it *V(C)*, all of the following seem to be true to how philosophers commonly use the method of cases.

[1] *C* can be an actual case, but does not have to be, and typically is not. Instead, *C* is typically a hypothetical case and the narrative details of *C* often involve counterfactuals and even violations of recognized scientific laws. In fact, philosophers are completely free to stipulate the material facts of *C*, although the more complicated the narrative details of *C* become, the less evidential weight philosophers are usually willing to give *V(C)*.

[2] Philosophers can appeal to *V(C)* as evidence without having to provide further evidence for *V(C)*. What is important about the practice is that *somehow or other* we have independent evidential entitlement to *V(C)*.

[3] Accordingly, the ways that philosophers can respond to *V(C)* are similar to the kinds of responses that are available to other kinds of evidence, and with corresponding costs. Philosophers can deny that *V(C)* is the right response to *C*, although at the risk of being cast as an outlier. Philosophers can also respond to *V(C)* by attempting to *explain it away*, according to some hypothesis about deviant ways the others might have arrived at *V(C)*.[6] Moreover, while philosophers are not allowed to affirm their own theories as reasons to undermine or override *V(C)*, they can continue to hold those theories in spite of *V(C)*, pleading such things as explanatory simplicity (Weatherson 2003).

There are surely more things to say here, but notice how far we can get in characterizing the practice without depending on, or disagreeing with, what different philosophers think about philosophical intuitions, or indeed on anything whatsoever about any particular kind of mental states whatsoever. For example, while the role and corresponding epistemic status of case verdicts is often understood to be similar to the role and corresponding status of perceptual evidence (Bealer 1998, Sosa 1998), such a view is not apparently mandated by the practice itself, and is a matter of ongoing debate (Williamson 2007, Cappelen 2012, and Deutsch 2015).[7] And this all makes perfect sense. If anything about how philosophers commonly use the method of cases told us anything interesting about whether the method required philosophical intuitions or any other special cognitive machinery, then we should expect to find evidence for this in our practices, themselves, and we simply do not. (For example, if our practices were to mandate that proper intuitions be emitted by well-functioning pineal glands, we would expect to find philosophers

[6] See Chapter 6 for further discussion.

[7] We will take up this debate in Chapter 2. But for now, we will just note that it is clear that Cappelen's proposed diagnostic characteristics for the presence of intuitions in a text are simply beside the point in terms of characterizing the method of cases as it is practiced. For Cappelen, intuitions possess both a characteristic phenomenology and a special *sui generis* indubitable epistemic status, and they are to be grounded entirely in terms of conceptual or linguistic competence (Cappelen 2012, 112–3). Obviously, nothing in the practice as we have described it involves any of those notions. That intuitions *sensu* Cappelen may fail to appear in a selected text simply has no bearing on the question of whether the method of cases is deployed in that text.

including recent MRI scans in their submissions to journals.) All that seems to be needed, in order to account for the shape of our practices, is that our case verdicts are adequately independent from our philosophical theories, such that the former may constitute evidence for the latter, that they be somewhat constrained by our attentional capacities and limitations, in order to explain the evidential penalty that can accrue with excessive length or detail, but that they be not overly constrained by our beliefs about what is or is not actually the case or even nomologically possible, in order to fit the fairly unlimited degree of permissible stipulation. Perhaps these verdicts are somehow subserved or implemented by some special kind of mental state or states (Nado 2014), but nothing about this seems to be built into the practice itself.

To be clear, we are not any sort of "intuition deniers" (Nado 2016) or anything like that. What's more, there are surely lots of good historical reasons that the current debate about the method of cases has centered on philosophical intuitions, or at least talk of "philosophical intuitions," such as the long tradition in epistemology of framing its inquiries in terms of the skeptical concerns about putative sources of evidence (with "philosophical intuition" naming one such source to be investigated), or the willingness of respected figures such as Kurt Gödel to posit faculties of intuition at least for mathematical knowledge (for further discussion, see Parsons 1995), or the rise to prominence of the term in Chomskyan linguistics (for further discussion, see Andow 2015). Our point is just that, from the perspective of philosophical practice itself, it seems that the method of cases doesn't really have much to say about the nature of any underlying, implementing, and/or enabling mental states one way or the other. So far as the method is concerned, our capacity to reach verdicts about philosophical cases is mostly a black box, or perhaps even a suite of black boxes.

Armchair Philosophy

So, our focus going forward will be on the method of cases rather than on any particular view about the kinds of mental states that are involved when we think about philosophical cases. More than just the method in general, though, we will be targeting here a particular way of deploying that method, namely, as a piece of *armchair* philosophy. Philosophers often elide the difference between the two methodological notions of the method of cases, on the one hand, and armchair philosophy, on the other, but one of the main points of the book will be to insist on this distinction. Very roughly, we will take "method of cases" to pick out a particular kind of evidence commonly used in analytic philosophy, while "armchair philosophy" describes a set of further resources that may—or, rather, may *not*—be recruited as a part of whatever methods one is using. Williamson (2007,

1) helpfully provides a nice gloss of how philosophers often talk about philosophical methods, including the method of cases:

> The traditional methods of philosophy are armchair ones: they consist of thinking, without any special interaction with the world beyond the chair, such as measurement, observation or experiment would involve.

On this rather standard way of thinking and talking about philosophical practice, the method of cases as standardly practiced is an armchair method. To see what this means, we need to both reframe and unpack Williamson's characterization of armchair philosophy. To do this, it is important to again distinguish between the mental processes—the *thinking*—that goes on when philosophers advance and evaluate philosophical arguments and the moves that philosophers make as part of those arguments themselves. And, for all of the reasons that we have rehearsed already, we think that the focus should be on the latter rather than the former. What it means to do philosophy from the armchair, then, involves identifying what kinds of premises and inferences can standardly be appealed to in a philosophical argument without any special further argumentation or evidence needed to license them. We think that we can group these into two main categories, depending on whether they are supposed to be part of a general, non-specialist background body of information or whether they are more properly seen as part of our shared philosophical training and heritage:

Common general knowledge: common sense, which would include the sorts of information available from informal observation, both perceptual and intellectual, and the kinds of common background knowledge that would be available to people with college educations, including even some fairly sophisticated scientific knowledge provided that this scientific knowledge is sufficiently well established at this point, such as the general outlines of modern physics and evolutionary biology. As a rough rule of thumb: we have in mind here the kind of scientific results that don't require someone to offer any citations on their behalf when they appeal to them, or just by referencing a name (e.g., "Darwin") or a title (e.g., "Special Relativity") without more specific reference.

Common professional knowledge: the history of philosophy, including standard views about what positions and arguments were successful or unsuccessful and why; first-order logic plus some basics of modal logic, in fact probably all formal logic and a fair bit of mathematics are fair game, no matter how technically demanding; and our shared theories of argumentation including a fairly well-theorized set of norms both positive (e.g., select premises that plausibly will be granted by an opponent; make clear how your premises

> collectively necessitate your conclusion) and negative (e.g., don't argue in a circle, don't confuse use and mention), together with norms for responding to arguments (e.g., what constitutes a successful counterexample). Again, as a rough rule of thumb: we have in mind here the kinds of historical claims that can be made without much explanation (e.g., that Hume successfully established that it is hard to derive prescriptions from descriptions, or that logical positivism *prima facie* does not satisfy its own criteria of meaning), the kinds of arguments that are acceptable (e.g., that a fairly compact deductive proof from a tolerably small set of premises remains the gold standard for how to get from premises to a conclusion in a philosophical argument), and so on.

There is surely more to be said about armchair methods, but this is enough to give us a better sense of what we think that the last couple of decades' debate about the method of cases is all about. It is a debate about a way of doing philosophy where philosophers take themselves to be licensed to appeal evidentially to verdicts about real or imagined cases without any sort of independent argument for those verdicts, *and* it is about a way of doing philosophy that philosophers think that they can pursue without any special sorts of empirical inquiry; basically, ordinary observation plus all of the logic you want, perhaps together with some set of accepted scientific background knowledge and history of philosophy. What we find ourselves engaged in is, in other words, not really a debate about the method of cases, per se, but an *armchair-bound version* of the method of cases, or what we will call the *armchair method of cases*, for short. Why does the restriction to the armchair matter? It will take us the whole book to really answer this question. For right now, we can sketch an answer using a common methodological metaphor. Philosophers who use the method of cases are committed to the idea that they have an independent evidential entitlement to case verdicts about real and imagined cases. This evidential entitlement is fallible, and this has led philosophers to realize, as part of their practice with this method, the importance of sorting out the good philosophical wheat from the problematic verdictive chaff. We can see this kind of project in play in multiple places in philosophical practice, from dialectical attempts to explain away unwanted case verdicts to various attempts to distinguish genuine philosophical intuitions from other sorts of presumptively wayward cognitions. The problem is that any armchair-bound version of this project is one that, at a minimum, forswears the ongoing help of the sciences in detecting problematic and worrisome case verdicts, and so we will argue that for all the good and valuable knowledge that counts as "armchair knowledge" by even a fair expansion of the Williamsonian gloss, an armchair-bound version of the method of cases will nonetheless include painfully little that can help with the hard work of the philosophical threshing.

2 Experimental Philosophy

Before we can say more about how we think that philosophers should start to think about the current debate about the method of cases, and why, we need to say more about how this debate got started and how much of it turns on the role, indeed the *necessary* role, of scientific methods in the kind of project that we just sketched. The current debate about the method of cases can be traced to the development of *experimental philosophy*, or *X-phi*, for short. We like to gloss the spirit of the X-phi movement as a movement where philosophers recognize that many of our philosophical inquiries, even many of the armchair ones, have empirical presuppositions and commitments—and take responsibility for seeing which of those presuppositions and commitments hold up under proper scrutiny. As such, X-phi includes a wide range of questions and methods for addressing them. Our interest here, however, is on one particular strand of experimental philosophy research, what we might call "the psychology of philosophy." One of the most prominent varieties (though, again, not at all the only variety) of X-phi involves the study of the way that people think about philosophical scenarios, including how they make the kinds of verdicts relevant to the method of cases. Some of this work has been meant simply to help philosophers learn more about just *what* people do, in fact, think about these kinds of cases, and so to help philosophers answer the kinds of epistemological, metaphysical, and ethical questions that Gettier, Frankfurt, Thomson, and others were interested in helping us answer when they developed their famous cases. But some of this work has been meant to shed light on what kinds of things shape *how* people think about these kinds of cases, and so to help philosophers think carefully about an approach to philosophical questions that involves using what we think about philosophical cases, or case verdicts, to try to answer those questions in the first place. And this kind of work, shedding light on what kinds of things shape how people think about philosophical cases, has given rise to the current debate about the method of cases, and is what we will focus on in this book.

Before we turn our attention to what this work has shown us about the method of cases, it is worth saying something about the way that we've just described experimental philosophy. The reason is that this way of laying out this way of thinking about X-phi might seem, at first glance, to be at odds with the way that Joshua Knobe (2016) popularly describes it, according to which X-phi is to be subsumed under the aegis of cognitive science. First impressions can be deceiving, and so it is important to see that these two ways of thinking about X-phi are not at odds with one another. In fact, we agree with Knobe that X-phi is cognitive science; we just don't think that is *all* that it is. Learning more about what people think about the kinds of philosophical cases used in epistemology, metaphysics, ethics, and so on, as well as the kinds of things that shape how people think about these kinds of cases, should certainly be interesting to cognitive scientists. But it should

also be interesting to epistemologists, metaphysicians, ethicists, and other philosophers. If it is not, then X-phi has failed to live up to its own advance billing, which was both to help philosophers answer the kinds of questions that we are interested in answering and to help philosophers better understand the kinds of methods that we can use to try to answer those kinds of questions. Experimental philosophers are scientists *and* philosophers, and experimental philosophy is both cognitive science and philosophy.

Since we just mentioned the way that experimental philosophy has been billed, it is worth saying something here about its origin story, especially since this origin story will help explain its crucial importance to the kind of project that we discussed in the previous section. One of the lessons of a century or so of empirical psychology has been that our minds are much more prone to error than we might have antecedently thought and, moreover, more prone to kinds of errors that can only be brought into view by painstaking empirical investigation. We find this picture of the human mind in psychological results about substantial and unexpected failures of first-person knowledge and the foibles of intuitive judgment, often in the face of significant reflection, and of our self-understanding regarding such basic faculties as perception or memory, and even among experts with a great deal of training and experience in their domains of expertise (see e.g., Tverksy and Kahneman 1974, Nisbett and Wilson 1977, Dawes et al. 1989, Haidt 2001, Wilson 2004, Loftus 2005, Simons and Ambinder 2005, Stanovich and West 2008, and for one classic discussion, see Stich 1990). This picture appears not only in first-order results within psychology, but even more so within the evolving methods that psychologists use, going back at least to the failure of introspectionist approaches and then over the last century or so with such key innovations as controlling for order effects, experimenter bias, and demographic variation. One of the central lessons, in fact, of the evolution and development of empirical psychology is that our minds *often don't even know what they don't know about themselves*, and this very much includes the sorts of factors that can shape our judgments across time and influence them at a specific time. And this is true even though we can sometimes spot some of the epistemically deleterious effects that play havoc on our thoughts. The fact that we can spot *some* of them *some* of the time does not give us much hope that we can spot *all* or even *many* of them *most* of the time. Thus, to tell the developmental story of experimental psychology is really in the same breath to tell the story of the limits and limitations of armchair psychology.

The intellectual genealogy of experimental philosophy, and its challenge to armchair philosophy, can in turn be mapped as a confluence of two streams of thought. The first is the picture canvassed in the preceding paragraph of the limitations of armchair psychology within experimental psychology, which then increasingly flows into philosophy through the philosophy of mind and interdisciplinary cognitive science; the second is a current, steadily strengthening within philosophy itself, of dissatisfaction with the method of cases. This second current dates back,

at least, to concerns that Benson Mates (1958) raised about how these methods were used in ordinary language philosophy, and then to a series of broad, and broadly naturalistic, attacks on intuition in philosophy (Cummins 1998, Kornblith 1998, and Stich 1998; see also Goldman and Pust 1998 for a critical but nonetheless more conciliatory naturalistic approach to intuitional methods). These sweeping criticisms of armchair philosophical methods, however, turned out to require really strong epistemological premises; so strong, in fact, that these early challenges turned out to be relatively easy to turn back or avoid (Bealer 1998, Sosa 1998).[8] Experimental philosophy, then, initially took root in an awareness within some corners of philosophy that our armchair philosophical methods may well be more susceptible to error than had been previously thought and, in particular, susceptible to errors that might be difficult, or even impossible, to detect and root out without substantial empirical investigation. The sense of security that philosophers have had in armchair philosophical methods, it seemed, was based in no small part on an armchair psychological theory of how well unaided human cognition can come to map its own contours, both its treasures *and* its pitfalls—a theory that experimental philosophers came to realize had been thoroughly discredited by developments within scientific psychology. And one promise of experimental philosophy was that we could use experimental methods to get a better view of where those sources of error might be, and thereby do better in our philosophical inquiries as a result.

If this is a founding promise in experimental philosophy's origin story, then how far have we gone in the meantime toward fulfilling it? Over the past two decades or so, two of the important things that we have learned are that case verdicts have an unexpected and unpredictable liability to be both *heterogeneous* and *unstable*. Different groups of people can reach systematically different verdicts about the same cases, often departing from established philosophical consensus, and the same people reach systematically different verdicts about the same cases when those cases are presented in different ways. Edouard Machery (2017; see especially Chapter 2) has recently provided an admirably comprehensive overview of this empirical work, and we see no need to recapitulate it here. Instead, let's focus here on two prominent examples, just as illustrations of these two different classes of X-phi findings.

Heterogeneity and the Method of Cases

One of the most famous examples of how the method of cases is used in philosophy is Saul Kripke's (1980) famous "Gödel case." Kripke begins by describing a person

[8] We will discuss this difficulty with finding an adequate epistemological premise at greater length in Chapter 4.

who claims that Gödel is the person who provided the incompleteness of arithmetic. This person knows enough mathematics to be able to give an independent account of the incompleteness theorem, but all that this person knows about Gödel is that he was the person who proved this theorem. With this background established, Kripke asks readers to consider the following hypothetical case:

> Suppose that Gödel was not in fact the author of this theorem. A man named 'Schmidt', whose body was found in Vienna under mysterious circumstances many years ago, actually did the work in question. His friend Gödel somehow got a hold of the manuscript and it was thereafter attributed to Gödel. (pp. 83–84)

According to the descriptivist theory of reference, when our fictional protagonist says that Gödel is the man who proved the incompleteness of arithmetic, he is referring to Schmidt because it is Schmidt who uniquely satisfies the description associated with the name 'Gödel' (namely, the person who proved the incompleteness of arithmetic). Yet, Kripke says that it doesn't seem like our fictional protagonist is referring to Schmidt; instead, it seems like the person is referring to Gödel even though some of his beliefs about Gödel have turned out to be false. And Kripke thinks that it is a strike against the descriptivist theory that it is at odds with what we naturally seem to think about this hypothetical case. Lots of other philosophers have agreed. The fact that the descriptivist theory of reference tells us that the protagonist in Kripke's story is referring to Schmidt when it seems natural to say that the protagonist is referring to Gödel suggests that it is not the case that names refer to whatever satisfies the descriptions that competent users associate with them. What's more, the causal-historical theory of reference successfully explains why it seems natural to say that the protagonist is referring to Gödel: the protagonist's use of the name is appropriately linked to how the name was first used—namely, to refer to Gödel. And Kripke thinks that the fact that the causal-historical theory of reference can explain why it seems natural to say that the protagonist is referring to Gödel suggests that reference is fixed by the initial act of naming; and later uses of names refer by being causally linked to these initial ("baptismal" or "grounding") acts.

Inspired by work by Richard Nisbett on cross-cultural cognitive diversity (Nisbett et al. 2001, Nisbett 2003), in one of the earliest papers in experimental philosophy, Edouard Machery, Ron Mallon, Shaun Nichols, and Stephen Stich (Machery et al. 2004) decided to see whether judgments about the reference of proper names vary across culture. Nisbett had shown that East Asians are inclined to make classification judgments on the basis of similarity and Westerners are inclined to focus on causation when making classification judgments, Machery, Mallon, Nichols, and Stich hypothesized that East Asians would be more likely to be descriptivists about the reference of proper names than Westerners. To test this

hypothesis, they presented participants in Hong Kong and the United States with cases similar to Kripke's Gödel case:

> Suppose that John has learned in college that Gödel is the man who proved an important mathematical theorem, called the incompleteness of arithmetic. John is quite good at mathematics and he can give an accurate statement of the incompleteness theorem, which he attributes to Gödel as the discoverer. But this is the only thing that he has heard about Gödel. Now suppose that Gödel was not the author of this theorem. A man called "Schmidt," whose body was found in Vienna under mysterious circumstances many years ago, actually did the work in question. His friend Gödel somehow got a hold of the manuscript and claimed credit for the work, which was thereafter attributed to Gödel. Thus he has been known as the man who proved the incompleteness of arithmetic. Most people who have heard the name "Gödel" are like John; the claim that Gödel discovered the incompleteness theorem is the only thing that they have ever heard about Gödel.
>
> Ivy is a high school student in Hong Kong. In her astronomy class she was taught that Tsu Ch'ung Chih was the man who first determined the precise time of the summer and winter solstices. But, like all her classmates, this is the only thing she has heard about Tsu Ch'ung Chih. Now suppose that Tsu Ch'ung Chih did not really make this discovery. He stole it from an astronomer who died soon after making the discovery. But the theft remained entirely undetected and Tsu Ch'ung Chih became famous for the discovery of the precise times of the solstices. Many people are like Ivy; the claim that Tsu Ch'ung Chih determined the solstice times is the only thing they have ever heard about him.

As predicted, they found that Chinese participants were more likely to make descriptivist judgments about both the Gödel case and the Tsu Ch'ung Chih case than American participants; in fact, *most* American participants gave causal-historical or "Kripkean" responses and *most* Chinese participants gave descriptivist responses. This work has been replicated extensively (Machery et al. 2009, Machery et al. 2010, Beebe and Undercoffer 2015, Machery et al. 2015, Sytsma et al. 2015, and Beebe and Undercoffer 2016), and suggests that our verdicts about at least some philosophical cases depend as much on who we are as it does on what the cases are supposed to show us.[9]

[9] We would note however that whether these observed differences should really be explained in terms of differing theories of reference is an open matter for debate. van Dongen, Colombo, Romero, and Sprenger (2021) perform an interesting meta-analysis of this empirical work, and Machery (2020) has an extensive literature review of the growing body of empirical work on reference. We will return to this work in Chapter 6, where we will see that recent empirical work suggests that the longstanding philosophical disagreement about the nature of reference rests on a *mistake*.

Instability and the Method of Cases

This fact about us, that what we think about some philosophical cases depends as much on who we are as it does on what the cases are supposed to show us, is just one of the surprising things that experimental philosophy has helped teach us about ourselves. Another is that it turns out that what we think about philosophical cases depends not only on who we are, but also on the ways in which we are asked to think about those cases. This is often called a *framing effect*. Tversky and Kahneman (1981) provide the classic example of how framing effects influence how people think about hypothetical cases.[10] They asked people to compare the following two cases:

> Imagine that the U.S. is preparing for an outbreak of an unusual Asian disease that is expected to kill 600 people. Two alternative programs to fight the disease, A and B, have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows: If program A is adopted, 200 people will be saved. If program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved. Which of the two programs would you favor?
>
> Imagine that the U.S. is preparing for an outbreak of an unusual Asian disease that is expected to kill 600 people. Two alternative programs to fight the disease, C and D, have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows: If program C is adopted, 400 people will die. If program D is adopted, there is a 1/3 probability that nobody will die, and a 2/3 probability that 600 people will die. Which of the two programs would you favor?

A little reflection shows that program A is the same as program C, and program B is the same as program D. And, yet, because the programs in each pair are described in different ways, Tversky and Kahneman found that, while most people who are asked to choose between programs A and B favor A, most people who are asked to choose between programs C and D favor D. They suggest that what is going on is that people are risk averse when the programs are described in positive terms, but risk seeking when the programs are described in negative terms. Whatever the merits of this explanation, what is important to note is that people's evaluation of the two programs seems to be influenced by the way that those programs are *framed*.

The kind of framing effect that we just described is one kind of framing effect, and is sometimes called a "content-based" framing effect. While content-based

[10] The classic philosophical discussion of this work in connection to ethical intuitions is Tamara Horowitz's (1998) paper "Philosophical intuitions and psychological theory."

framing effects have gotten a lot of attention, especially among psychologists and other social and cognitive scientists, experimental philosophers have also been interested in another kind of framing effect—one that involves the context in which cases are presented. Let's call this a "context-based" framing effect. What people think about hypothetical cases is influenced by this kind of framing effect when those evaluations depend in important ways on how they are asked to think about the case or on what other things they have been asked recently to think about.

A prominent example of this kind of framing effect in philosophy comes from recent experimental work on different versions of the famous trolley problem that we described above, work that was inspired by a groundbreaking study conducted by Lewis Petrinovich and Patricia O'Neill (1996). Here's one of the studies, which focuses on Judith Thomson's famous "Loop case" (Liao et al. 2012). Here is the case in question:

> Loop: A runaway trolley is headed toward five innocent people who are on the track and who will be killed unless something is done. Abigail can push a button, which will redirect the trolley onto a second track, where this is an innocent bystander. The runaway trolley would be stopped by hitting the innocent bystander, thereby saving the five but killing the innocent bystander. The second track loops back toward the five people. Hence, if it were not the case that the trolley would hit the innocent bystander and grin to a halt, the trolley would go around and kill the five people.

And here are the two cases that were used to frame the loop case:

> Standard: A runaway trolley is headed toward five innocent people who are on the track and who will be killed unless something is done. Abigail can push a button, which will redirect the trolley onto a second track, saving the five people. However, on this second track is an innocent bystander, who will be killed if the trolley is turned back onto this track.

> Push: A runaway trolley is headed toward five innocent people who are on the track and who will be killed unless something is done. Abigail can push a button, which will activate a moveable platform that will move an innocent bystander in front of the trolley. The runaway trolley would be stopped by hitting the innocent bystander, thereby saving the five but killing the innocent bystander.

Some participants were asked to think about the loop case after first being asked to think about the standard case, while other people were asked to think about the loop case after first being asked to think about the push case. And it turns out that participants were more likely to think that pushing the button in the loop case was morally permissible when they had first been asked to think about the standard case than when they had first been asked to think about the push case. Since what

seems to matter is not how the cases are described, it seems that what is causing the framing effect is instead the context in which the case is being evaluated, and so this study suggests that what we think about at least some philosophical cases depends as much on the contexts in which we are asked to think about them as it does on what the cases are supposed to show us. Similar kinds of content- and context-based framing effects have been found using different versions of the trolley problem (Valdesolo and DeSteno 2006, Nadelhoffer and Feltz 2008, Uhlmann et al. 2009, Cikara et al. 2010, Schwitzgebel and Cushman 2012, Wiegmann et al. 2012, Pastötter et al. 2013, Costa et al. 2014, Wiegmann and Waldman 2014, and Schwitzgebel and Cushman 2015).

Heterogeneity, Instability, and the Limits of Armchair Methods

What is particularly interesting about heterogeneity and instability, and the reason why we have focused on them here, is that these are precisely the kinds of things that experimental philosophers, drawing on the lessons learned during the development of experimental psychology, suspected would be difficult or even impossible to detect and root out without substantial empirical investigation. Cognitive diversity in philosophy, after all, is likely to be obscured by our natural tendency to sample those who are like us and amplified by selection effects within the profession. A famous story told by the film critic Pauline Kael about the 1972 presidential election helps illustrate how biased sampling and selection effects can work in professional contexts: "I live in a rather special world. I only know one person who voted for Nixon. Where they are I don't know. They are outside my ken. But sometimes when I'm in a theater I can feel them." In the context of presidential elections, we end up getting good feedback about the extent to which the political views represented in our local communities may diverge from the larger body politic, and Kael was probably much more sensitive than the average person to the ways in which communities diverge from one another owing, no doubt, to the time she spent in theaters and to her acute attention to popular cinema. In contrast, philosophers turn out to be rather more at sea about whether or not they are on the same wavelength when it comes to the kinds of standard—and, standardly, at least a little bit weird—kinds of cases philosophers use when we use the method of cases (see, for example, Nahmias et al. 2006 on attitudes toward compatibilism and incompatibilism).

It is of the utmost significance for philosophical methodology that both of these sorts of effects, by their nature, lie hidden out of view from someone perched upon an armchair. Cognitive diversity is often obscured by sampling bias and other selection effects. Philosophers, at least within the Anglophone analytic community, are a problematically homogenous bunch, and individual philosophers resting comfortably in their armchairs, even talking to other philosophers in other armchairs,

are not likely to gain much of a sense of what diversity may be out there.[11] And framing effects that we talked about in connection with the trolley problem are often obscured by the fact that, for the most part, they operate on us unconsciously, and we almost never confront the same case in multiple contexts or framings in close enough succession to notice. This makes them invisible to introspection and unlikely to be revealed from a first-person perspective, or by casual observation from a third-person one. In fact, these kinds of effects are precisely the kinds of effects that it took experimental psychology to uncover in the first place, and that make it clear to psychologists the advantages that experimental psychology has over armchair psychology. All of this helps highlight not only the fact that heterogeneity and instability lie beyond the power of our armchair methods to detect and correct, but also something else often overlooked in debates about the armchair method of cases: the sources of error that experimental methods reveal may also be ones that such methods can help overcome. Good experimental methods can control for many such effects directly, with studies designed to sample deliberately across a broad range of participants where cases are presented in different orders to different participants, as is fairly standard scientific practice, and where statistical methods are used to help pick out real effects from spurious ones. Moreover, where hypotheses arise about less easily controlled factors, the potential influence of such factors can be studied directly.

Much of the so-called negative program in experimental philosophy, from its earliest days, has focused on investigating heterogeneity and instability. They were the sorts of hypotheses that drove much of the research in the first generation of this methodologically critical side of experimental philosophy.[12] But in the burgeoning literature of experimental results, another kind of problematic finding has emerged, not one that anyone was setting out to look for, but which has nonetheless been repeatedly found. In a way it is a special sub-case of heterogeneity, but we think it is worth considering in its own right: namely, that many cases that the philosophical literature take to be decisively instances of *F* or of not-*F*, the folk have overwhelmingly found to be . . . kinda sorta *F*-ish, kinda sorta not-*F*-ish. For example, in our own early work with Stacey Swain (Swain et al. 2008), we reported an order effect on Keith Lehrer's (1990) "Truetemp" case. Truetemp is a perfectly reliable estimator of local temperatures, but he is unaware of his reliability and has no reason to think he has such a capacity. Can Truetemp properly be said to *know* the temperature when he estimates it? Lehrer's reported verdict, much shared in the epistemological community, is that Truetemp does not know, and this verdict

[11] We discuss this in more detail in Chapter 3, when we talk about the "numerator problem." For a comprehensive empirical study on diversity in philosophy, see Schwitzgebel et al. (2021).

[12] See, for example, the "four hypotheses" of the earliest such paper, Weinberg et al. (2001): that philosophical case verdicts might vary by culture; by socio-economic status; by degree of philosophical training; or by order of presentation of cases. In essence, three heterogeneity hypotheses and one instability one.

has been taken as one to be accommodated both by internalists like Lehrer and the externalists, like Goldman, against whom the case was originally deployed. We reported a modest order effect, in which different groups of subjects had somewhat different mean evaluations of the case depending on what sort of epistemological scenario they might have been presented with beforehand. That order effect we reported has had a mixed record of successful (Wright 2010) and unsuccessful (Ziółkowski 2021, Ziółkowski et al. 2023) replication, and we are not putting any argumentative weight on that claimed instability effect here. Rather, one interesting result from all these studies of Lehrer's case is that the folk have pretty consistently found Truetemp to be *not* a clear failure of knowledge, but at most a very intermediate one. In our own early study, we found rates of attribution of knowledge on a 5-point Likert scale ranging from 2.4 (tilted about half a point toward "is not knowledge") to 3.6 (tilted about half a point toward "is knowledge").[13] In Jennifer Cole Wright's (2010) replication of our study, she did not use a Likert scale, but instead reported percentages of attributions. Of the three orders she looked at, only one of them showed a strong inclination against Truetemp's estimate counting as knowledge (26% attribution of knowledge), but the other two orders had either only a modest lean that way (40% attribution), or indeed a modest lean the other way (55% attribution). In Adrian Ziółkowski's (2021) failed replication of our study, all of the presentation orders yielded inconclusive verdicts, with rates of attribution ranging on a 5-point Likert scale from 3.0 to 3.4 across the three studies that he conducted. All in all, these results do not paint a picture of the Truetemp case being the resounding case where the protagonist lacks knowledge that the epistemological literature has taken it to be, but rather a fairly *inconclusive* case in terms of knowledge. There are many other examples of this sort of mediocrity of case verdicts in the literature, for example Bergenholtz et al. (2021) on "fake barn" cases. Not much has been made of these verdicts of mediocrity to date, because it has generally not been hypothesized and investigated in its own right, just uncovered as a byproduct of work chasing down other sorts of claims.

Now, there is nothing in principle to preclude inconclusive case verdicts serving as evidence, especially for theories that might predict just such intermediate responses at just such places.[14] The problem is not that intermediate cases exist, so much as philosophical practice seems to be poorly attuned to them. This *inconclusiveness* of many case verdicts is inconsistent with the way that they are almost always deployed in armchair practice, to count strongly for one theory and against others. The armchair method of cases apparently has inadequate safeguards

[13] The distributions were normal, so the responses were more or less clustered around these means.

[14] Another way of trying to bring these kinds of case verdicts back into play would be to argue that the inconclusive nature of folk verdicts about philosophical cases is simply a sign that there's something wrong with the folk. We will discuss this kind of move, which we have called elsewhere (Weinberg et al. 2010) the expertise defense, in Chapter 5, and make our case there as to how ineffective such a defense is in facing down the experimentalist challenge. In a nutshell: no such expertise claim can even remotely be secured from the methodologically impoverished seated position.

against mistaking mushy verdicts for solid ones.[15] This sort of error may arise because of a general divergence between philosophers and the folk as to how they render different case verdicts. Yet it might also arise as a problem in the practices, without being a problem in the practitioners themselves: while one might generally expect that what is received in the literature reflects a consensus verdict among philosophers at large, this need not be so. It may also be that some sub-literatures on occasion get captured by practitioners who agree among themselves, but are divergent from the profession on the whole. From the armchair, we can often discern accurately what the consensus around a given case is in a literature, but it is impossible to see when such a consensus has grown around a verdict that is inconclusive among the folk; and if so, just what the nature of the divergence may be.

3 Experimental Philosophy and the Method of Cases

We hope that the broadest contours of a problem for the method of cases have now come into view. On the one hand, the armchair method of cases presupposes an entitlement to deploy a wide range of case verdicts as philosophical evidence, even, and indeed most typically, in the absence of any specific empirical evidence that would license our reliance on such verdicts. On the other hand, cognitive psychology in general, and experimental philosophy in particular, have shown that the human capacity to render these sorts of case verdicts suffers from some substantial shortcomings that we cannot expect to be generally detectable or correctable from the restricted epistemic vantage of the armchair. Let's stipulate the term *robust verdicts* to pick out those verdicts which are not subject to evidentiary defects like heterogeneity, instability, and inconclusiveness.[16] Such verdicts are strong and univocal, and while that is no guarantee of veracity, of course, they are at least the right kind of thing to plug into the method of cases. With this bit of terminology in place, we can now succinctly present the *experimental challenge to armchair philosophy* in broad outline: the armchair method of cases presupposes that the case verdicts as appealed to in philosophical practice are robust, whereas we have good empirical evidence that verdicts fail to be robust more often than philosophers have thought; and that a big reason for this unpleasant surprise is that robustness failures can be very hard, impossible even, to detect from the armchair.[17]

[15] We have an extended discussion of "stakes effects" in epistemology as perhaps making a similar sort of "mountains out of molehills" error in Chapter 7.

[16] For our purposes here, we will define robustness strictly in terms of these three categories. But should X-phi reveal *yet further classes* of armchair-undetectable sources of error, we would want "robust" reconstrued to include them.

[17] We thus end up in a position surprisingly similar to that of Ichikawa (2016)—surprising, because he has long been a defender of the armchair from experimentalist arguments. We all seem to agree that the X-phi worries do not depend on any particular view of what intuitions are, and indeed do not require taking any psychological states to constitute the evidence that philosophers use. The central issue,

The armchair practitioner is thus challenged to offer a defense of their decision, given this situation, for remaining in the armchair instead of embracing more experimental tools. The project for this book is to spell out the terms of the experimental challenge and what it should mean for philosophical practice that those terms cannot be met.

In order to set out our own novel construction of the experimental challenge, we first need to dismantle the now standard way of thinking about the challenge. As things stand, the current debate about the method of cases has largely focused on whether or not the method is *reliable*. This might seem like a natural way to frame any debate about philosophical methodology, for at least two reasons. The first reason is that reliability is a common and well-understood epistemological concept and, as Jonathan Ichikawa (2016, 155) notes, "No discipline that affords a central evidential role to an unreliable source is in good standing." The second reason is that the fact that different groups of people reach systematically different verdicts about the same cases, often departing from established philosophical consensus, and the same people reach systematically different verdicts about the same cases when those cases are presented in different ways seems quite straightforwardly to suggest that the method of cases is *unreliable*. Edouard Machery (2017) is the latest to advance this kind of argument against the method of cases, arguing that since unreliability reasonably counts strongly against a method's evidentiary fitness, philosophers should stop relying on the method of cases as evidence. The rub, according to Machery, is that since the method of cases plays a seemingly ineliminable role in helping philosophers answer the kinds of questions that interest us, this means abandoning not only the method of cases but also many of the questions that we have come to associate with philosophy.

Let's call this the *standard normative framework* for the debate about the method of cases. In Chapter 2, we will address two concerns that have been raised recently about this way of thinking about the relationship between analytic and experimental philosophy. The first concern, which has been raised by Max Deutsch (2010, 2015) and Herman Cappelen (2012), is that the standard normative framework fundamentally misrepresents the nature of analytic philosophical practice: the method of cases does not depend essentially on what we think about philosophical cases, but is instead an *argumentative* practice where philosophers argue for verdicts about philosophical cases. The second concern, which has been raised by Joshua Knobe (2019, 2021, forthcoming), is that the standard normative framework fundamentally misrepresents what experimental philosophy has

on our shared understanding, is that the way we form our verdicts can go awry, and in ways that may need scientific help to suss out. One key difference is that he uses a personal epistemic framework to express this possible relevance of X-phi to philosophy, in terms of one's having, as an individual philosopher, "possible reason to doubt one's own ability to respond rationally to the available evidence." In contrast, we would put the point in public and methodological terms, something like our having 'possible reason to doubt how safe the method of cases is from errors'.

revealed about what people think about philosophical cases: experimental philosophy has revealed that a *surprising* number of verdicts about philosophical cases turn out to be neither heterogeneous nor unstable. We will argue that both attempts to raise concerns about the standard normative framework from the perspective of analytic and experimental philosophical practices fail precisely because neither is sufficiently attentive to the particulars of those practices. The moral of this story, we will contend, is that philosophers simply cannot continue to argue about philosophical methodology without really looking at philosophical methodology, as it is actually practiced.

Having defended the standard normative framework from these concerns, we will nevertheless argue in Chapter 3 that philosophers need to move past it, and to shift away from thinking about the method of cases in *epistemic* terms. Not only does it turn out to be really hard to determine whether the method of cases is reliable or unreliable, and for reasons that are built right into the very concepts of *reliability* and *unreliability*, but when it comes to methodological questions, how reliable our belief-forming practices have to be in order to count as reliable enough turns out to be relative to specific domains and purposes. Even moderately reliable methods, including ones that operate only slightly above chance, can be put to good use, at least in principle and very often in practice, provided that we are *careful* about how we use them. And we will argue, following Nado (2016), that the mistake that most experimental philosophers have made is to look for something about the method of cases that is root-and-branch epistemologically problematic.[18] We think that experimental philosophy's dirty little secret is that the method of cases is probably not in such a bad way that it cannot *sometimes* be put to good use *somewhere* and *somehow*.

Because we think that experimental philosophy does not put us in the position to make strong negative epistemological generalizations about the method of cases, in Chapter 4 we will argue that the best way forward in the current metaphilosophical debate about the method of cases is a wholesale abandonment of the urge to think about the method of cases in anything like the conceptual framework of epistemic normativity. A better way to think about the experimental challenge is within a conceptual framework we will call *methodological rationality*. What we have in mind is a special case of practical rationality that can be applied to different research methods, including the method of cases. It is not a *sui generis* form of rationality, but instead a way of thinking about philosophy in terms of both its practical goals and what resources and methods we should adopt in order to maximize the chances that we can achieve these goals. To give a quick preview of what this change in focus will mean for debates about the method of cases, let's

[18] Again to be clear, we are only considering a proper subset of experimental philosophers on the whole, namely, those that have been interested in applying X-phi tools to the method of cases. We ask the reader to keep this quantifier restriction in mind throughout the rest of the book.

return to the way that we characterized the method of cases earlier in this chapter. Remember that we said that the armchair method of cases is usually understood to be a method that philosophers can use without any special sorts of empirical inquiry; basically, ordinary observation plus all of the logic you want, perhaps together with some set of accepted scientific background knowledge. It is standardly practiced as an armchair method, and as Williamson astutely points out right before his characterization of armchair methods quoted above, "Every armchair pursuit raises the question of whether its methods are adequate to its aims" (2007, 1). Changing the debate from the conceptual framework of epistemic normativity to the framework of methodological rationality puts the kind of question that Williamson says is raised by our armchair methods front and center: indeed, *are* our armchair methods adequate to the aims we have for using them? To address that crucial question, we don't need to ask anything like, what is so wrong with the method of cases that we shouldn't use it at all? Trying to answer *that* question turns out to be a mug's game. Instead, we can ask, rather more fruitfully: what methodological benefits would we get from leaving the armchair; and would those benefits outweigh whatever costs come with doing so? After all, it is a direct entailment of the way that Williamson construes armchair philosophy that leaving the armchair is a methodologically richer position than remaining comfortably seated. We thus reframe the debate about the method of cases as a cost-benefit analysis of continuing to use this method from the armchair versus making use of the rich methodological resources that come along with leaving the armchair behind.

While the costs of extracting ourselves from our collective armchairs will need to be explored, particularly in terms of what that would mean for both the philosophical community at large and for individual philosophers, we will argue that the benefits will be well worth the costs. To start, in Chapter 5 we will argue that there are significant costs associated with staying in the armchair, due to armchair philosophy's painful insensitivities to many risks of error in our case verdicts. As we discussed earlier in this chapter, we have growing evidence that case verdicts may fail to be robust more often than antecedently expected. And the limited methodological resources that are available to philosophers when we restrict ourselves to armchair philosophical methods are simply not the kinds of resources that are needed to detect and correct for this kind of evidential heterogeneity, instability, and/or inconclusiveness.

As a short illustration of the methodological costs of "unknown unknown" sources of error, consider how Sosa (2007b) tries to caution that the experimental challenger needs to grant to the method of cases (discussed in terms of "intuitions") at least the same resources that we grant ourselves in our ordinary epistemic lives. He contends that

> the effects of priming, framing, and other such contextual factors will affect the epistemic status of intuition in general, only in the sort of way that they affect the

> epistemic status of perceptual observations in general. One would think that the ways of preserving the epistemic importance of perception in the face of such effects on perceptual judgments would be analogously available for the preservation of the epistemic importance of intuition in the face of such effects on intuitive judgments. The upshot is that we have to be *careful* in how we use intuition, not that intuition is useless. (105)

But as we have argued elsewhere (Alexander and Weinberg 2007), *it can only pay to be careful when we know what it means to be careful*. And here Sosa's analogy to sense perception is helpful, although not for the reasons that he thinks that it is, as it brings out just how impoverished armchair philosophy is in this regard. Our ordinary practices provide substantial (though incomplete!) resources to us to inform our use of sense perception, and which help us know when and how to use sense perception in a responsible way. We have a pretty good understanding of when sense perception goes wrong, something that is reflected in our perpetual practices and reinforced by a communal scientific understanding of the mechanisms responsible for our perceptual judgments. This prevents worries about problematic kinds of perceptual error from giving rise to global concerns about the use of perceptual methods. The problem is that armchair philosophical resources just don't put us in nearly the same position with respect to verdicts about the kinds of real and imagined cases that are interesting to philosophers. In a sense, they just can't tell us what it would mean to be careful in concrete, practical, *methodological* terms. As we will argue in Chapter 5, our armchair practices demonstrate an adequate understanding neither of where armchair-indetectable errors might arise nor of what positive steps might be taken to search for, correct for, or otherwise protect our inquiries from them.

Our main contention, then, will be that analytic philosophy is currently in a state of *methodological irrationality*. When we restrict ourselves to the methodological resources that have traditionally been associated with armchair philosophy, we leave ourselves unreasonably vulnerable to error. What's more, we will also argue in Chapter 6 that the benefits associated with adding experimental tools and methods to the set of methodological resources that are available to us when we use the method of cases are not limited to satisfying conditions of methodological rationality, or simply helping us render the method of cases less vulnerable to error; adding experimental tools and methods offers a potential boon to philosophical progress, something that adds significant philosophical benefits to that side of the ledger. We will conclude, then, in Chapter 7, by arguing that all of this means that we should make a substantial investment into the kinds of error-management methods that are available when we expand our methodological toolbox to include the kinds of scientific methods that are stock-in-trade for experimental philosophers.

There's obviously much more to be said—in fact, a whole book's worth of things! Before we get underway, we want to underscore that our book is offered in a spirit

of profound optimism for analytic philosophy. To that end, let us contrast our way of thinking about what experimental philosophy means for analytic philosophy with the way that Edouard Machery (2017) thinks about its significance and implications. Machery argues for a *bounded* conception of philosophy's ambitions, where we abandon the method of cases and with it many of the philosophical questions that we have rightly come to associate with philosophy over the course of its history. We think that philosophers should opt instead for an *unbounded* conception of philosophical methodology. Taking philosophical questions seriously means getting serious about what kinds of methods we should use to answer them, and this means focusing not only on questions of epistemic normativity but instead on questions of methodological rationality, and learning what methodological advantages experimental philosophy can bring to the method of cases. And so long as philosophers are willing to engage seriously with the ways in which experimental methods can complement more traditional, armchair philosophical methods, we can hope for real progress. This is what *standing up for philosophy* is all about: giving ourselves a richer set of methodological resources that can be used to answer the kinds of philosophical questions that philosophers have been interested in asking all along.

2
Metaphilosophy and Meta-Methodology

Before we can start to talk meaningfully about the implications that the experimental challenge has for the method of cases, it is important for us to address two concerns that have been raised recently about the way that we have been talking both about the method of cases and about experimental philosophy. Some analytic philosophers, most notably Max Deutsch (2010, 2015) and Herman Cappelen (2012), have argued that the way that we have been talking, here and elsewhere, about the method of cases reveals a fundamental misunderstanding about the nature of philosophical practice. Here's how Deutsch puts this concern:

> Analytic philosophy is chock-full of hypothetical examples and thought experiments, of course, but analytic philosophers *argue* for their claims about what is or is not true in these cases and thought experiments. It is these arguments, *not* intuitions, that are, and should be, treated as evidence for the claims. (2015, xv)

The basic concern appears to be this. The way that we have been talking about the method of cases focuses too much attention on something that turns out not to matter all that much to philosophy and philosophical practice, at least from an evidential point of view, namely, our *intuitions* about these cases.[1] What's worse, the way that we have been talking about the method of cases focuses too little attention on what really does matter to philosophy and philosophical practice, namely, philosophical arguments. If this is right, then analytic philosophers can sidestep the kind of experimental challenge to the method of cases that we rehearsed in Chapter 1 and that we will develop in much more detail in the next few chapters. If Deutsch and Cappelen are right, then nothing that we can learn from studying philosophical cognition is relevant to philosophical methodology since our philosophical methods involve argument, not intuitions.

Joshua Knobe (2019, 2021, and 2023) has raised a second concern about the way that we have been talking about the method of cases and experimental philosophy, arguing that the way that we have been talking about them overlooks the fact that experimental philosophy has revealed that a *surprising* number of verdicts about

[1] This is one of the places that we mentioned in Chapter 1 where recent debates about philosophical methodology are dominated by "intuition"-talk, and so we are going to adopt this way of talking in this chapter despite the reasons that we gave in Chapter 1 for wanting to avoid this language as much as possible. Here, we think that the benefits of adopting the *lingua franca* outweigh the costs.

Standing Up for Philosophy. Jonathan M. Weinberg and Joshua Alexander, Oxford University Press.
 DOI: 10.1093/9780191947704.003.0002

philosophical cases turn out to be neither heterogeneous nor unstable. If this is right, then analytic philosophers don't have to cut the experimental challenge off at the pass, they can cut it off at the knees. As Knobe writes,

> If people's intuitions are surprisingly stable, then the whole debate about the implications of instability is moot. As we noted at the outset, there has been an enormous amount of research focused on the question: "If we learn that people's intuitions are unstable, what should we conclude about the use of intuitions in philosophy?" Such research has shown impressive levels of sophistication and ingenuity, but if people's intuitions are not in fact unstable, this is simply not the question we face. (2021, 69–70)

In this chapter, we will try to show that both concerns miss the point. It is perhaps unsurprising that we think that they do, given the skin we have in this game. What we think is surprising, and our reason for spending a chapter talking about these concerns, is why they miss the point and what this tells us about the way that we should approach questions about philosophical methodology. In a nutshell, they fail because they take their eyes off the most important part of the game: the methodological practices themselves.

1 Arguments All the Way Down?

Let's start with the suggestion that case verdicts do not actually figure in our philosophical methods.[2] Are Deutsch and Cappelen right that all, or even most, of the evidential work being done in philosophy is being done by philosophical arguments rather than by what philosophers think about philosophical cases? Deutsch and Cappelen advance this view by proposing reinterpretations of many famous "case-based" philosophical books and papers according to which the case verdicts that seem to play a central role in these papers depend on philosophical arguments. One problem with this strategy is that their "no-intuition" interpretations of these cases have struck many readers, including us, as controversial at best.[3] Here we will focus on two of their most prominent case studies, Gettier (1963) and Lehrer (1990), and argue that their method of, in essence, squinting very hard at the target texts has serious shortcomings and leaves the debate potentially stalemated. In

[2] We are only focusing on the "arguments not intuitions" line of thought in these authors. See Chapter 1, Note 6 for why we are not engaging here with other aspects of Cappelen's arguments in particular.

[3] David Colaço and Edouard Machery (2017) make a similar observation in their review of Deutsch's book; Avner Baz also does so quite forcefully in his (2017). Cappelen's argument has other problems, most glaringly that he operationalizes "intuition" in an implausibly strong way that makes it almost trivial that philosophers do not use intuitions *sensu* Cappelen. For discussion, see Weinberg (2014a).

short, because the relevant textual passages are fairly brief, evidence for any interpretation of what's going on in those passages runs out quickly, and there's something inherently unresolvable in debates about philosophers' mental states from decades past. After raising worries about the way that they try to establish their interpretation of these cases, we will show what we can learn about these cases by paying closer attention to the broader contours of ordinary philosophical practice. We think that this practice-oriented methodological approach can bring substantially more evidence to bear, and so we hope to be able to achieve the first-order result of pushing back hard against the no-intuition interpretation and the second-order result of underscoring the importance of attending to practices.

Let's start with Gettier, since Deutsch offers such a sustained and thought-through attempt to interpret this classic text in terms of arguments, and not intuitions about the cases. Most people who have written about Gettier's thought experiments, and there have been lots and lots of them, have agreed both that it sure seems an awful lot like Smith lacks knowledge, and that it sure seems an awful lot like Gettier thinks that the fact that it seems an awful lot like Smith lacks knowledge counts as evidence that sometimes our justified true beliefs do not count as knowledge. Deutsch disagrees, arguing that although Gettier's case against the view that knowledge is justified true belief centers on the claim that Smith's justified true beliefs do not count as knowledge, Gettier does not rely on this claim because it is intuitive, but instead because he has a good argument that Smith's justified true beliefs do not count as knowledge. Where is the argument? Deutsch points our attention to the very end of the first of Gettier's two cases:

> But it is equally clear that Smith does not *know* that (e) is true; for (e) is true in virtue of the number of coins in Smith's pocket, while Smith does not know how many coins are in Smith's pocket, and bases his belief in (e) on a count of the coins in Jones's pocket, whom he falsely believes to be the man who will get the job.

According to Deutsch, all of the argumentative work is being done by this sentence. On Deutsch's reading of this sentence, Gettier's case that knowledge is not simply justified true belief depends on how we interpret Gettier's use of the word 'for'. Deutsch thinks that Gettier is using the word 'for' as a premise indicator, and this makes it seem rather clear to him that Gettier intends what follows to count as evidence that Smith does not know that (e).

We have substantial worries about Deutsch's interpretation of the passage. But before we get into those, let's suppose for the sake of argument that he is right and Gettier is indeed arguing to, rather than from, the case verdict in question. It is important to see that this concession is not enough to really put all that much pressure on the idea that experimental philosophy is relevant to how we should think about philosophical methodology. Here's why. If Getter is making an argument here, then *he is not making very much of one* (Mallon 2017 presses this point). Deutsch (2016)

seems willing to acknowledge this point, and suggests that much of the argumentative work being done when philosophers use thought experiments might be happening off-stage, so to speak, in the heads of the philosophers who are constructing the cases and the readers who are consuming them. This includes what evidence philosophers, both *producers* and *consumers* (to borrow some terminology from Landes 2023) of thought experiments, think they have for the premises being appealed in those arguments, explicitly or otherwise. But the social and cognitive sciences give us lots of reasons to worry that our evaluations of what evidence we have for our beliefs depend a lot on those beliefs, and so what we happen to believe, and what kinds of things influence what we happen to believe, is always going to be relevant, although perhaps in a less direct way than is sometimes suggested. This is more than enough to get experimental philosophy, and the kinds of worries that experimental philosophers like us have wanted to raise and address about how philosophers use the method of cases, up and running. Experimental philosophy is relevant to understanding philosophical methodology because, *whatever it turns out to be*, on Deutsch's account, that philosophers are tacitly invoking in their arguments, we know that the tacitness of those invocations in and of itself suggests that they may well be susceptible to unwanted biases and influences, which are just the sort of thing that experimental tools and methods are needed in order to root out. That philosophers would, of course, want to avoid any such errors caused by our unconscious cognition provides plenty of reason to recognize the relevance of experimental philosophy.

Putting this aside, Deutsch's interpretation of what is going on in the passage quoted earlier is far from mandatory. There are other ways to read how Gettier is using the word 'for' in that passage, readings that are equally plausible and that challenge the idea that Gettier intends what follows to count as evidence that Smith does not know that (e). According to one alternative reading, Gettier is using the word 'for' to signal the start of an *explanation* for why Smith does not know that (e).[4] The difference between these two readings is subtle, to be sure, but it is also important. According to Deutsch's reading, Gettier starts with something like the idea that knowledge is incompatible with certain kinds of epistemic luck and, apparently expecting such a generalization to be acceptable as a shared premise, argues from there that Smith does not know that (e). According to the more common reading, however, Gettier starts with the intuitive claim that Smith does not know that (e) and explains why Smith does not know that (e) by highlighting the role that a specific kind of epistemic luck plays in his cases.

[4] Colaço and Machery (2017) also press this point. Cappelen and Deutsch are aware of this distinction, and Deutsch attempts to address the alternative reading in his (2016). We don't find the reasons that he gives in favor of the argumentative reading of the relevant text over the explanatory reading compelling for reasons that we get into below. See Conte (2022) for a similar defense of the explanatory construal of argumentative language in the totally different philosophical domain of political philosophy.

William Ramsey (2019) offers another alternative reading, according to which thought experiments like Gettier's are intended to provoke an intuition in readers, and the supporting text is offered in order to help *cue* the intuition that is intended. As he explains it,

> thought experiments provide a prompt for an intended psychological reaction. But such a prompt works only if the audience grasps the relevant details of the scenario. To ensure this, philosophers often reiterate salient elements so that the intended target of persuasion realizes the key specifics. This is what Gettier is doing in this sentence [i.e., the one quoted above]. His example is complicated, so he simply reemphasizes the most pertinent aspects for anyone who didn't follow along, and thereby failed to have the intended reaction. In this way, Gettier is acting like someone trying to prod the recall of a crucial memory by relaying noteworthy aspects of whatever occurrence the memory is about. (pp. 92–93)

Similarly, consider trying to get students to see an alternate percept in a bistable illusion case, for example, they are stuck seeing the duck-rabbit only as a rabbit. You might say to them things like, "you can see it as a duck, for the rabbit's ears could be like the two halves of a duck's bill, and the duck is facing left instead of right," and in doing so, you would in no way thereby be *arguing* for the anatine visual interpretation over the leporine. You are aiding the student in seeing it for themselves, and coming to their own unargued-for perceptual realization. Which of these readings of Gettier is right? If we just stare hard at this short stretch of text, this question seems, at best, highly debatable. The high plausibility of the two alternative readings that we have sketched here precludes any easy inference to the claim that, while Gettier treats the claim that Smith does not know that (e) as evidence, he only does so because he has an argument for it.[5] This sort of dispute is going to be hopelessly mootable so long as philosophers just engage in hermeneutic squinting, but as we will argue shortly, once we consider the much broader set of evidence about philosophical practices, the question becomes much less moot.

Looking beyond this brief passage itself, what really should matter for these debates is not so much what Gettier took himself to be doing, or even what anyone else has taken Gettier to be doing, but what the published literature tells us *about philosophical methodology as practiced in the last half century and counting*. Not that Gettier's own self-understanding is irrelevant, by any means. But if most epistemologists did not read him as offering an argument, but rather an intuition, and

[5] We do not intend this list of three possible construals of the argument-seeming discourse—inferential, explanatory, and attentional—to be exhaustive. And indeed, there are likely more than two alternative readings—Bengson (2013) catalogs many further supporting but not supplanting uses of argumentative language in the context of the method of cases. See also Chalmers (2014), Weatherson (2014), and Brown (2017). For example, see Angelucci (2022) for a construal in dialectical, not evidential, terms.

thereupon engaged in a practice that trades in intuitions, and if that practice is the one that continues today, then that clearly will be more salient to how we understand the method of cases, and philosophical methodology more generally, than Gettier's three pages from when gas cost $0.30 a gallon. Deutsch does make an attempt at some of this in his (2016), but we fear his examples do not ultimately work in his favor. At best they suggest one possible way to read some philosophers who, to our eyes, are theorizing about an anti-luck condition, a condition that seems to us best understood as independently motivated by case verdicts, rather than arguing from an anti-luck condition to establish those case verdicts. What's more, it doesn't seem to us that Deutsch's readings of these philosophers and their work are well supported by the total evidence available.

So, for example, Deutsch talks about Alvin Goldman's discussion of Gettier cases in his famous (1967) article "A causal theory of knowing." But in the first passage that Deutsch quotes, where Goldman is describing his strategy for the article, Goldman writes "Michael Clark, for example, points to the fact that *q* is false and suggests that as the reason why Smith cannot be said to know *p* . . . I shall make another hypothesis *to account for the fact* that Smith cannot be said to know *p*, and I shall generalize this into a new analysis of '*S* knows *p*'" (Goldman 1967, 358; emphasis added). It seems clear to us that Goldman's plan is to start from the position that Smith does not know that *p*, explain why Smith's justified true belief that *p* doesn't count as knowledge, and then use this explanation to build a new theory of knowledge. It does not, in other words, seem to us that Goldman is planning to use his theory of knowledge to *argue* that Smith doesn't know that *p*, which is what Deutsch needs Goldman to be planning to do in order for the strategy that he is taking in his (2016) paper to work.[6]

While Deutsch's strategy in his (2016) paper doesn't seem to us to be persuasive, let's grant it to him as a defensive move, that is, that some epistemologists may have read Gettier the way that he suggests. The problem is that, even when we grant this to Deutsch, it is just not hard to find other, very highly regarded epistemologists who patently cannot be assimilated to this argument interpretation. Here we will just focus on two. First, Robert Shope (2002, 30–31), a highly engaged participant in the 'S knows that P' debates and widely considered *the* expert on Gettierology, writes in a handbook piece on the Gettier problem that

[6] It is actually unsurprising that Goldman seems to start with the intuition that Smith doesn't know that *p* given all of the rather important and influential work (for example, Goldman and Pust 1998, Goldman 2007) he has done on the role that intuitions play in philosophical methodology, and so it is somewhat odd that Deutsch uses him as a prime example here. In fact, the first sentence of Goldman (2017), which in fairness appeared after Deutsch's article, makes it clear precisely what Goldman's take is on this debate: "Reactions to Gettier's (1963) paper demonstrated the powerful role of intuitions in philosophical methodology."

> Gettier offered no diagnosis of these examples and no formula for constructing further examples that he was prepared to regard as of the same type. But as other philosophers proceeded to offer additional examples that they regarded as importantly similar to one or another of Gettier's, the technical label, 'Gettier-type example', sprang into use.

This does not sound like someone who took Gettier, or the crowd of respondents to Gettier, to have agreed on an anti-luck premise that they were just further investigating! It sounds, instead, very much like someone who found himself, and took his colleagues to similarly find themselves, with an unargued-for evidential access to the verdicts about Gettier's and a great many and sundry similar cases.

Second, in the later editions (e.g., (1989)) of Roderick Chisholm's *Theory of Knowledge*, he chooses to present only the second of Gettier's cases without the disputably argumentative material at the end of the first case. He describes Gettier as having "noted"—not a verb of argumentation or inference—that the situation in the case "is counter to the traditional theory of knowledge" (91), and in his own presentation of various cases, Chisholm's typical style is to present the case sharply and efficiently, and issue a verdict, with nothing at all offered in support even of the minimal sort that Deutsch proposes to find in Gettier's own paper. Since Chisholm was arguably the most prominent and influential analytic epistemologist of the day, it seems unlikely that his way of approaching Gettier was a highly aberrant one-off. For what it's worth, what would be very helpful at this point would be for some historians of analytic philosophy to offer a scholarly take on just how mid-century epistemologists understood their methods; we suspect that some combination of ordinary language philosophy and Chisholm's own "particularism" was highly conducive to what would now look like an intuition-based methodology, but that is basically just conjecture on our part.[7]

So much for Deutsch's interpretation of Gettier's famous thought experiments; let's turn our attention now to Herman Cappelen's analysis of Lehrer's famous *Truetemp* case. The case involves a person who is a perfectly reliable estimator of local temperatures, but who is unaware of his reliability in this capacity, and has no reason whatsoever to think he has such a capacity. Can this person be said properly to *know* the temperature when he estimates it? Lehrer, on almost everyone's reading, invites readers to share the intuition that none of Truetemp's true beliefs count as knowledge. And this intuition is used to make a lot of trouble for *externalist* theories of knowledge, since at least some of those theories, especially *reliabilist* ones, would seem to predict that Truetemp does know. On this common reading, Lehrer is appealing to intuitions about the case and arguing from them

[7] Keep in mind that Gettier was himself a student of Norman Malcolm.

toward a rejection of externalism. Cappelen construes matters the other way around, using a bit of Lehrer's text:

> The primary argument [that Truetemp doesn't know] goes something like this: "More than the possession of correct information is required for knowledge. One must have some way of knowing the information is correct" ([Lehrer], 188). Since Mr. Truetemp has no way of knowing that the information is correct, he does not know. (p. 168)

As we saw with Deutsch's interpretation of Gettier, just attending closely to a few sentences of text can make it hard to determine precisely what is the right way to read those sentences. And Lehrer's own words do invite some confusion. For example, after describing the case, Lehrer almost immediately goes on to say: "the sort of causal, nomological, statistical, or counterfactual relationships required by externalism may all be present. *Does he know that the temperature is 104 degrees when the thought occurs to him while strolling in Pima Canyon? He has no idea why the thought occurred to him or that such thoughts are almost always correct. He doesn't, consequently, know that the temperature is 104 degrees when that thought occurs to him*" (187; emphasis added by Cappelen). Now, Cappelen looks at this and declares, "I don't know how to read this other than as Lehrer giving an argument in favor of a certain answer to the question" (170).[8] *Pace* Cappelen, it seems to us entirely clear how to read aspects of this passage in terms of the explanatory interpretation that we rehearsed above. That 'consequently' could be inferential, as Cappelen suggests. But it could also easily be explanatory, if we take Lehrer to be laying out how his internalist hypothesis correctly predicts the intuited verdict. (It is admittedly harder to see how to apply the third, cueing interpretation to 'consequently' here.) Since Cappelen says nothing more about his interpretive blockage, let's just leave the point here: the method of peering deep into the soul of a couple of quoted passages is not up to resolving such interpretive disputes.

But, once again, when we step back from the vignettes themselves and their immediate textual surroundings, there is plenty more evidence to be found. First, there are *other* parts of Lehrer's text that strongly indicate that he cannot be arguing in the principle-to-case direction, but the other way around. Lehrer is very explicitly embedding his discussion in a dialectic with externalists, and he means to be giving reasons that those theorists could themselves recognize and take to be problematic for their externalist views. It directly follows that we just can't take as any sort of shared premise between Lehrer and externalists that having "no idea why the thought occurred to him or that such thoughts are almost always correct" will entail a lack of knowledge. Indeed, at the start of the relevant chapter,

[8] Deutsch concurs: "It is very difficult to see how anyone could read this passage as anything other than an attempt to provide reasons for the judgment" that Truetemp doesn't know (112).

Lehrer says this plainly: "The central tenet of externalism is that some relationship to the external world accounting for the truth of our belief suffices to convert true belief to knowledge *without our having any idea of that relationship*" (177; emphasis added). What Cappelen takes to be a premise is what Lehrer clearly states is the very issue at stake with his opponents. *Contra* Cappelen, once we bring into view the broader argumentative context in which the case is embedded, we don't see how to read this as Lehrer giving an argument in favor of a certain answer to the question. As Landes (2020, 19) observes, in an excellent elaboration of this argument:

> Lehrer takes himself to have access to what knowledge is independently of the externalist theories of knowledge he is discussing. For this access to justify his rejection of externalism, it cannot be derived from an internalist theory of knowledge. Otherwise, because this passage is used to argue against the externalist views of Armstrong, Nozick, Goldman, Dretske, and others, Lehrer would be flagrantly begging the question against his opponents.[9]

Moreover, it's hard for the no-intuitions, arguing-from-principles-to-cases approach to make sense of how Lehrer introduces this whole line of argument (p. 186):

> The general opacity problem with externalism can be seen most graphically by considering the analogy proposed by Armstrong. He suggested that the right model of knowledge is a thermometer. The relationship between the reading on a thermometer and the temperature of the object illustrates the theories mentioned above. Suppose that the thermometer is an accurate one and that it records a temperature of 104 degrees for some oil it is used to measure . . . The problem with the analogy is that the thermometer is obviously ignorant of the temperature it records. The question is—why?

There's plainly no argument to the verdict there; just the flat, if utterly plausible, declaration of its obvious truth. And the question then sure seems to be an explanatory one: we have an intuition, and we can now try to make some philosophical headway by explaining it. We take this to be some further evidence of Lehrer's approach to his Truetemp case as well, since Lehrer presents the Truetemp case as in essence an elaboration of the literal thermometer case.

As we also saw with Deutsch's interpretation of Gettier, since we are investigating a widely deployed philosophical methodology, what the original author of a case

[9] Although we don't want to pursue this line of argument here, Landes makes an excellent case that similar concerns can be raised for any attempt to construct a non-question-begging version of the alleged Gettier argument in the Smith case.

may have thought about it is much less important than how the profession picked up and ran with it from there.[10] And here's a useful observation about the analytic epistemology literature since the 1980s–1990s: Lehrer's Truetemp case often shows up in discussions juxtaposed with a highly similar set of cases due to Laurence BonJour (1985), about a set of clairvoyants who nonetheless lack any clue that they are, in fact, clairvoyant, and whose clairvoyant beliefs are thus in much the same epistemic boat as Truetemp's meteorological ones. Indeed, the differences between the cases are considered cosmetic enough that epistemologists on the whole have taken Lehrer's Truetemp and BonJour's clairvoyants to be *more or less indistinguishable*. For just one recent example, here's Mylan Engel (2022, 41):

> Internal reasons reliabilism also yields the intuitively correct verdict on both BonJour's Norman [one of the clairvoyants] case and Lehrer's TrueTemp case. Why? Because both Norman and TrueTemp lack internal reasons for their respective beliefs.

Some quick poking about on Google Scholar will demonstrate how widespread is this general assimilation of the two cases. We thus propose that we have good *prima facie* reason to take it that the analytic epistemological community at large takes these cases to work in fundamentally the same way. Most importantly here: if either is a principle-to-case argument, then the other ought to be interpreted as such as well, and vice versa.

But, unlike the potential interpretive difficulties induced during a hyper close inspection of Gettier's use of the word 'for', or Lehrer's 'consequently', BonJour is exceedingly clear on how the order of argument is meant to go:

> But it seems intuitively clear nevertheless that this is not a case of . . . knowledge: Samantha [another clairvoyant] is being thoroughly irrational and irresponsible in disregarding cogent evidence that the President is not in New York City on the basis of a clairvoyant power which she has no reason at all to think that she possesses; and this irrationality is not somehow canceled by the fact that she happens to be right. Thus, I submit, Samantha's irrationality and irresponsibility prevent her belief from being epistemically justified.

Now, the second half of that quoted paragraph, especially that 'thus', certainly can make it look like BonJour is arguing for the conclusion that the clairvoyant doesn't know on the basis of premises about irrationality of a certain sort precluding

[10] Cappelen claims in an aside, "It's important to focus on the original text, not the argument as it is idealized in the later literature" (169). For a scholar of the original author, surely that is so. For methodologists and metaphilosophers, however, Cappelen's advice is perfectly wrong. See also Colaço and Machery (2017, 410–411) on the importance of considering the full historical context of arguments and methods in use.

knowledge. It can seem we're in another argument versus explanation bind. But then see where BonJour goes to immediately from there:

> The case and others like it [i.e., all the clairvoyant cases] suggest the need for a further condition to supplement Armstrong's original one: not only must it be true that there is a law-like connection between a person's belief and the state of affairs that makes it true, such that given the belief, the state of affairs cannot fail to obtain, but it must also be true that the person in question does not possess cogent reasons for thinking that the belief in question is false. For, as this case seems to show, the possession of such reasons renders the acceptance of the belief irrational in a way that cannot be overridden by a purely externalist justification. (p. 60)

The second quoted paragraph reveals why the "arguing for the verdict" interpretation could not actually be a good reading of the first paragraph: the case verdict itself is what is doing the argumentative work, as a premise, in revealing to us the relevant claim about irrationality and knowledge, which is itself the conclusion in turn.

That the cases are leading the way is further substantiated by BonJour's own account of his methodology, when he writes earlier in that paper that the radical nature of externalism makes it hard to find much common ground at the level of general principles about knowledge or justification:

> The problem, however, is that this very radicalism has the effect of insulating the externalist from any very direct refutation: any attempt at such a refutation is almost certain to appeal to premises that a thoroughgoing externalist would not accept.
>
> My solution to this threatened impasse will be to proceed on an intuitive level as far as possible. By considering a series of examples, I shall attempt to exhibit as clearly as possible the fundamental intuition about epistemic rationality that externalism seems to violate. Although this intuition may not constitute a conclusive objection to the view, it is enough, I believe, to shift the burden of proof decisively to the externalist. (p. 56)

This description of his own methodological approach should substantially undermine any attempt to read BonJour, or Lehrer with him, as offering arguments from the epistemic generalities to the cases. He needs the case verdicts themselves to be doing the argumentative work.

And BonJour is not being idiosyncratic here. Our reading of the literature is that it is very common, when philosophers deploy arguments with case verdicts in them, that these case verdicts are methodologically basic in the manner that we have suggested here, that is, with no further argumentation on their behalf. There

are many other examples where philosophers seem to say that their capacity to offer any further defense of a key premise has run aground. So, for example, Jerry Fodor (1997, 154) writes,

> As with most of the metaphysical claims one comes across these days, the one that I just made relies for its warrant on a blatant appeal to modal intuitions. But I think the modal intuitions that I'm mongering are pretty clearly the right ones to have. If you don't share mine, perhaps you need to have yours looked at.

In similar spirit, David Lewis (1996, 561) writes,

> I started with a puzzle: how can it be, when his conclusion is so silly, that the skeptic's argument is so irresistible? My Rule of Attention, and the version of the proviso that made that Rule trivial, were built to explain how the skeptic manages to sway us—why his argument seems irresistible, however temporarily. If you continue to find it eminently resistible in all contexts, you have no need of any such explanation. We just disagree about the explanandum phenomenon.[11]

Over and over again we find case verdicts being used in the way that we have described them being used. They are very frequently not, in fact, argued for from independently convincing and available premises. Nor is there any ready-to-hand generally available evidence to offer on their behalf, which can be presupposed to be antecedently shared, of a sort that we can find intelligible in terms of observation of testimony from experts, or, for that matter, from things like the deliverance of well-calibrated instruments, the products of scientific consensus, and so on. Philosophers seem on the whole fairly happy to use them as evidence, even in the absence of any further evidence on their behalf. The profession seems to be in a state of substantial consensus, then, that we have some sort of cognitive grasp or other on these claims.[12] And this consensus also extends pretty well to the basic contours of this capacity, for example, that it includes, but is not at all limited to, a pretty open-ended range of hypothetical cases with stipulated facts, including perhaps nomologically impossible ones. What the profession seems also to be in a state of substantial dissensus about, however, is just what this grasp ultimately amounts to, as illustrated by the many-hued panoply of extant accounts of case verdicts that we rehearsed earlier in the chapter. But none of this changes what

[11] Although Lewis talks about arguments in this passage, it is clear in the article that he doesn't have in mind anything like what Deutsch has in mind when he claims that the method of cases involves arguments. Here's how Lewis describes what he has in mind: "The sceptical argument is nothing new or fancy. It is just this: it seems as if knowledge must be by definition infallible" (1996, 549). That sure looks a lot like an intuition and not an argument. We are open to being told otherwise by Lewis scholars, but it seems that in general such disagreements about the "explanandum phenomenon" are not further adjudicable by argument.

[12] For additional discussion, see Chapter 1 and Chapter 5.

we have said about the role that the method of cases is meant to play in philosophical practice nor the evidential weight given to what philosophers think about those cases.

We have been suggesting so far that the claims of "intuition deniers" (a term coined by Nado 2016b) come to grief when they confront broader readings of the texts in which thought experiments appear. A further problem arises when we consider not just our practices regarding thought experiments themselves, but also our more general practices involving how we are taught both to construct arguments and to critically engage with them on the whole. Construing many of these thought experiments as arguments in the way the intuition deniers suggest flies in the face of our methodological norms of argumentation more generally. So, for example, norms of argumentation in analytic philosophy tend toward the *hyperarticulation of premises, presuppositions, and inferential structure*. And there are good reasons for these kinds of explicitness-encouraging norms. Well-articulated arguments are, well, articulated, and thus allow us to focus attention on each of their attendant parts. We can evaluate each premise, and the ones that seem at all dubious can be scrutinized even more closely. And, when the premises are presented in such a way that the reasoning becomes as close to transparently valid (or at least cogent) as possible, then we can carefully evaluate the evidential relationship between premises and conclusion. (Notice, by way of underscoring all of this, that it is often considered a very substantial demerit when someone's argument turns out to be enthymematic.) The kinds of arguments that Deutsch and Cappelen have in mind, and which they think can be located in famous verdict-based philosophical arguments, fail to provide any of the methodological benefits that proper and explicit arguments do, and so are not worthy of any particular respect or deference. If philosophers are making arguments in their heads, but *don't actually give us their arguments*, then really all they're doing is asking us to take their word on it. That's strange enough. What's even stranger is that they'd be asking their philosophical opponents to just take their word on it, as well, which is the oddest kind of dialectical trick. Suppose instead that philosophers are not arguing, but are instead issuing case verdicts in accord with the consensus professional norms of the armchair method of cases, which include the presupposition that one's own intuitions will most likely be shared by one's interlocutors. In this case, the overall shape of their argumentative behavior makes much better sense. Of course, it is possible that when philosophers are making arguments in their heads, they have some reasonable expectation that their interlocutors will reconstruct those arguments in *their* heads, and so aren't asking anyone to take their word for it. The problem is that there are limits to any charitable reconstruction of someone else's argument, and so limits to how charitable we can reasonably expect other people to be when they engage with our work. And so a dangerous game is afoot when philosophers leave this much to their readers, as demonstrated by our previous discussion about the different ways that philosophers have interpreted what's going on in Gettier's

paper. There is a reason that we just don't do this. Anyone with any experience with referees #2 will know how unlikely it is that even their explicitly stated arguments will be uniformly construed charitably, let alone their unstated ones.

The way that Deutsch and Cappelen want us to think about philosophical thought experiments also fails to make sense of how we are properly taught to engage with one another's actual arguments. So, for example, the act of *explaining away* what we think about philosophical cases would constitute a dramatic violation of the principle of charity were we to think that this involves explaining away someone's argument. To see this, consider how John Hawthorne (2003) and Timothy Williamson (2005) attempt to explain away what we think about epistemic cases that involve the possibility that the protagonist is making some kind of mistake or another in terms of the influence that the *availability heuristic* has on how we think about philosophical cases (for critical discussion, see Nagel 2010). We will discuss both "explaining away" and recent debates about how best to understand what people think about these kinds of epistemic cases in Chapter 6. For right now, it is important to contrast how this kind of argumentative move seems perfectly good when it takes aim at philosophers' verdicts about philosophical *cases* with how poorly received would be any version of this move that targets philosophical *arguments* themselves.[13] When attributing a heuristic to someone's subpersonal cognition, the main constraint in practice is that it be at least somewhat independently motivated, though in practice this need not involve an appeal to the scientific literature. But when proposing a reconstruction of someone's argument, there are significant constraints of charity. "You have asserted that *p*, without further argumentation, but I wonder if you really have in mind a patently fallacious argument for *p* that you would never reflectively endorse?" is *not* a move in good order in our argumentative practices.

We don't take ourselves to have responded here to the whole range of approaches taken by the intuition deniers (for review and discussion of these approaches, see Nado 2016b, Weinberg 2016a, and Horvath 2022), or to have come close to a conclusive refutation of the "arguments-not-intuitions" approach. In fact, as we will discuss in Chapter 7, an important lesson that should be learned from thinking carefully about the ways that Deutsch and Cappelen defend the method of cases is that philosophers need to be considerably more straightforward than they have been about just what they are doing when they use thought experiments. But we do take ourselves to have shown that the "arguments-not-intuitions" approach has so far not attended adequately to key particulars of philosophical practice, not just in the way that verdicts about thought experiments get deployed, but also in how they

[13] We should distinguish our argument here from that of Climenhaga (2018). Climenhaga argues that the practice of requiring theorists to explain away problematic intuitions is *in and of itself* excellent evidence that intuitions are treated as evidence; and we applaud this argument. Our argument here, however, is about how the Deutsch/Cappelen approach cannot make sense of *the particular sorts of explanations* that typically get offered in this practice. Our thanks to an anonymous referee on this point.

get argued *about*, and more generally, with the norms for philosophical practice for argumentation itself.

2 Much to *Whose* Surprise?

As we noted above, the first attempt to respond to the experimental challenge means to do so by evading the challenge altogether. If intuitions do not matter much to analytic philosophers, then any concerns that experimental philosophers might raise about them don't speak to how philosophers use the method of cases. Joshua Knobe, in a set of recent and forthcoming papers, raises a different kind of worry, arguing that the experimental challenge rests on a different kind of mistake. According to him, the experimental challenge doesn't get philosophical practice wrong, it gets the empirical results wrong, or at least is not consistent with the growing body of empirical work that has emerged over the past decade or so, including work on what has, or has not, replicated. It turns out, according to Knobe, that philosophical case verdicts are "surprisingly stable"—that is, less heterogeneous than we might have first thought and less sensitive to irrelevant factors, as well. And so, the experimental challenge turns out to be no challenge at all.

While Knobe is attempting to cast doubt on the key empirical premise of the experimental challenge, it is important to be clear that he is not trying to endorse armchair philosophical methods—which, since Knobe is a noted champion of experimental philosophy, would hardly be expected. He has his own metaphilosophical fish to fry, turning on an interesting distinction between *instability* and *tension* in our philosophical intuitions. Knobe thinks that there's very little of the former sort of thing, but lots of the latter, and he advances a specific research program to investigate these kinds of tensions. Our interest here, however, is just with his assessment of the kinds of stability or instability that are relevant to the experimental challenge. As we shall see, though, his characterization of how much stability there is or is not depends on what perspective one takes on the question.

We should start with a surprising feature of Knobe's papers, which is that their conclusions are framed explicitly in terms of whether the degree of stability found in people's philosophical intuitions is itself *surprising*. The adverb "surprisingly" is right there in the title of both papers, and he's very clear about it in his text, for example, when he writes (2021, 1),

> The evidence now suggests that philosophical intuitions are surprisingly stable. Indeed, the available evidence suggests that philosophical intuitions are surprisingly stable across both demographic groups and situations.

Or, when he writes (2019, 33),

> I have been suggesting that one surprising finding coming out of the experimental philosophy literature is the shocking degree to which demographic factors do not impact people's philosophical intuitions.

Surprising *and* shocking!

In one sense, just about the least surprising thing that we can see in philosophy is the word "surprising" in a paper about experimental philosophy.[14] After all, experimental philosophers *love* to characterize their results in such terms, and for the very good reason that their results often are indeed not ones that we would have antecedently predicted. And after all, much of the point of experimental philosophy is that it can reveal facts about philosophical cognition that are not available from the armchair.[15] But standardly in such papers, we can parcel out the experimental results themselves from the further gloss or spin concerning the surprisingness of those results. The study as conducted, the data that is gathered, the statistical inferences from that data—when it goes well, all of that can maybe tell you something about what you were studying, but none of it can tell you whether it adds up to anything surprising. And by and large, in our experience, referees don't seem to worry too much about the surprisingness bit, since in these papers it's mostly just rhetoric.

In Knobe's recent arguments, however, it can't be just rhetoric. There is literally no claim to be defended here without *something* setting a threshold for the observed degree of stability being claimed. There's no question, after all, that there's some amount of demographic variation greater than zero and less than absolute. For that matter, we expect everyone in this debate would agree that there's substantially more non-variation than variation, that is, that *most* philosophical intuitions will be invariant to *most* dimensions of possible demographic variation. And so, Knobe's "surprising" modifier is a crucial part of creating a thesis that would be, even in principle, empirically evaluable. Or consider this claim: "At an empirical level, the key question is how to explain the surprising robustness of philosophical intuitions. One possible answer would be that the capacities underlying people's philosophical intuitions have an innate basis" (33). Knobe's claim is only intelligible with some sort of characterization of *just how much* robustness there is, a function played here by that "surprising." That is, without the "surprising" or some other threshold-setting locution to characterize just how much robustness has been observed, there just wouldn't be an *explanandum* there for an innateness hypothesis to serve as a potential *explanans*.

[14] Just by way of illustration, Google Scholar shows 1,540 hits for "experimental philosophy" in articles that were published in 2021, and just over half (777) of them include "surprising." In contrast, there are about 285,000 hits for "philosophy," and only about 16,200 include "surprising"—about 6%.

[15] This idea received some pushback in Dunaway et al. (2013), but see Liao's (2016) critical response.

But to define a thesis in terms of *p*'s being "surprising" presupposes, at a minimum, some sort of priors regarding *p*. The thesis would then be that the empirical evidence mandates some heavy-duty updating from those priors. For Knobe's purposes, where might these priors come from? Once we recognize that there are several distinct but equally legitimate ways of filling that in, it becomes clear that there are a number of claims that Knobe could be putting into play here—claims that might not all agree in truth-value. Our main contention here will be that, while his claims are at the very least defensible and debatable for some candidate priors, for the source of priors that would be most relevant to debates about methodology, Knobe is badly off-target regarding that state of play in the empirical literature.

The first, and most literal, place to look for priors, in order to determine whether or not philosophical case verdicts are surprisingly stable, would be at what priors experimental philosophers give to the stability of philosophical case verdicts. Now, we'll just speak for ourselves because it's hard to speak for everyone. And, indeed, the fact that it obviously makes no sense to speak for everyone here should raise some preliminary worries that this may not be a great way to proceed. (Maybe we need to do some X-phi on the X-phi experts?) Anyhow, just reporting our own priors, while some of the specific experimental results that have been published in recent years have been surprising, we would say that overall the *trend* is not one that we think falls at all outside our expected range, for two main reasons. First, our view before learning about these experimental results was that philosophers, including experimental ones, didn't know hardly anything at all about just *where* and *how much* of *what sorts* of variations might be found. So the range of plausible distributions of variation was antecedently very open. We take it that what the early experimental philosophy studies showed was not, and was never meant to be, "*look, we've shown that there's rampant variation all over the place, afflicting everyone everywhere all at once*," but something much more like, "*look, we've found a bit of variation, and it wasn't hard to find, and no one really has much of a clue yet just how much variation there is or isn't out there. Intuition-deployers beware!*" The early studies revealed our fairly drastic ignorance about intuitional variation by revealing several previously uninvestigated hypotheses to be live empirical possibilities.

Knobe seems to be coming from a different starting point here, for example, when he writes that because experimental philosophers were studying "intuitions about seemingly abstruse issues, such as the nature of the true self or whether the universe is governed by deterministic laws. There was every reason to expect that such intuitions would differ radically between demographic groups" (2019, 31). We're not sure that there's any useful way to argue about conflicting priors, and so we'll just underscore how far apart our priors are here. It seems to us that there was every reason to think that philosophical intuitions about cases involving these philosophical issues might *differ occasionally, perhaps frequently and perhaps*

systematically, and in unexpected ways between demographic groups. And we take this to be very much weaker than what Knobe reports as his own previous view.

Our second reason to take ourselves to be unsurprised here is that we just don't think that the experimental studies that Knobe musters in his two papers go very far toward removing that ignorance. Here's how Knobe characterizes his argument in the 2019 paper:

> I have been suggesting that one surprising finding coming out of the experimental philosophy literature is the shocking degree to which demographic factors do not impact people's philosophical intuitions. In support of this claim, I have cited 30 studies, by 91 different researchers, comprising a total sample size of 12,696 participants. Many of these results would be highly surprising even in isolation. Taken together, they are downright shocking.

With all due respect to Knobe, we invite the reader to pause for a moment to consider just how huge the Cartesian product is when looking at the space of possible philosophical cases and the different ways in which philosophical case verdicts about these cases might vary. We think it will then seem fairly obvious that thirty studies is just not very many at all, especially when it's not a random sample, and he himself acknowledges that there are other studies that do find demographic variation.[16] People could only be shocked if they expected gobsmackingly wanton demographic variation every which way but loose, but we must confess that that strikes us as an implausible take on the initial experimental philosophy results, even from the very heady early days where effects seemed fairly easy to find.

In short: we already didn't have any very specific expectations, and we don't take ourselves to have received much evidence at all to disconfirm what expectations we did have, in their louche vagueness. It seems to us that the burgeoning body of work has trimmed off the more extreme ends of the distribution of possibilities here. On the one hand, it does not seem likely at this time that, say, East Asian and Western communities have differences in their knowledge attribution so stark as to motivate positing a substantial form of epistemic relativism. That seemed a real possibility in the wake of the original Weinberg et al. (2001) study; it no longer seems especially plausible now, especially given the repeated failed attempts at replication (e.g., Kim and Yuan 2015, Seyedsayamdost 2015, Machery et al. 2017). But on the other hand, when there were just a handful of results in, it was just as possible that *none* of them could have held up under further scrutiny *and* that no further interesting variation results would surface. Despite these notable replication

[16] Stich and Machery (2022) argue that Knobe also *misreads* the evidence. Knobe (2023) offers a reply. We will not pursue this line of debate here ourselves since our arguments are intended to hold even if Stich and Machery are incorrect in this matter. Of course, if they are in fact right, then so much the worse for Knobe's case against the empirical premise. See also Machery (2023) for further discussion.

failures of some of those early findings, so many new results have emerged and continue to emerge that it is no longer a live possibility that there are just no meaningful variations here to be concerned with. (See, for discussion, Machery 2017, especially Ch. 2.)

Another place to try to source the priors for a surprisingness claim is from *the state of the scientific literature*. This seems a better claim for Knobe, in terms of getting a claim that may well be true. Perhaps we could construct a history of experimental philosophy and the psychology it tends to attend to, starting in something like the 1990s in big cultural group differences, with work by folks like Richard Nisbett, Kaipeng Peng, or Jonathan Haidt, and then later the research by Joseph Henrich and others on the pervasive differences between Western cognition and that of the rest of the world, and also the explosion of social priming results from John Bargh, Simone Schnall, and others, and the way that the heady first few years of experimental philosophy seemed to be riding that wave—but then observe, quite rightly, that much of the subsequent work on these issues has ranged from the deflationary to the outright debunking of much of the earlier work. It seems to us legitimate to claim that there has been a swing in the big psychological sciences pendulum here, and that the smaller trajectory of experimental philosophy has resonated with it. In that sense, then, the shifting of momentum from reporting all sorts of variation to reporting all sorts of failed replications of variation is certainly noteworthy, and may also legitimately count as "surprising."

Not only is the "direction of the literature" construal of "surprising" a good bet for making Knobe's claim come out true, but we would note that questions about the large-scale scientific picture of the nature of the human mind are a perennial interest for Knobe, as evidenced in papers like his 2010 BBS target article. So we think that this way of setting the threshold in Knobe's conclusion is a charitable one, having him making a defensible claim of legitimate relevance to varieties of scientific debates that Knobe has been interested in. We do not take any stance here as to whether that threshold is or isn't met.[17] Our main contention here will be, rather, that questions about the direction of the scientific literature will not serve to set the "surprise" bar *at a spot that will be of much relevance to debates about philosophical methodology*.

To consider how much stability should count as surprising for the purposes of debating philosophical methodology, we should extract the relevant priors from the methodological practices themselves. What sorts of variation, and to what extent, are our methods anticipating? Or rather, to avoid any problematic anthropomorphism, how much variation of what kinds are our practices well configured to detect, and have been appropriately buttressed to handle? As we will argue

[17] Though, we suppose that given our earlier point that we don't take the total state of evidence to add up to so very much, probably we would opt for "isn't," if we had to bet. We'd probably opt more for "suggestive" than "surprising." But that plays no role in our arguments here.

in Chapter 5, it seems to us the answer is: *approximately none whatsoever!* One quick way to see this, in anticipation of the longer discussion in that chapter, is to consider what our philosophical practices would look like if we were at least *trying* to handle demographic and situational variation from the armchair, and to see how there is nothing of the sort in our current practices with the method of cases.[18] For starters, if our philosophical practices were anticipating any nontrivial amount of demographic or situational variation, then we would at a minimum see some attention and discussion of such variation, especially in handbooks and textbooks, with debates in the literature about the nature of these variations and how to handle them.[19] For demographic variation, for example, we might expect to find a norm of disclosure about every author's location among various parameters: ethnicity, gender identity, native language, religious upbringing, and so on. That way, we could increase our chances of spotting such variation where it might potentially arise. And, for our methodological practices to monitor for and, ideally, control for context or order effects, we would perhaps expect to see the development of specifications for some sort of canonical conditions for using the method of cases. Manuals for philosophical methodology might specify: sit in a quiet room, clear your mind with at least three minutes of meditation, then look at the scenario, as printed with black ink on white paper in a clear, sans-serif font. (Maybe Arial rather than Times New Roman MT Std would become *the* font of philosophical thought experiments?) What you would *not* see in philosophical practice are hypothetical cases just dropped into papers wherever it nicely suits the flow of the argument. We leave it to the reader to speculate on what sorts of rules could be adopted for the consideration of cases on the fly, for example, in a colloquium Q&A, in order to screen off such effects, and to observe that, of course, no such rules are even remotely in force, or even under consideration.

We are being fanciful here, but to a purpose: to highlight just how very little our philosophical practices with the method of cases are prepared for really any instability whatsoever. Contrast this with our extensive resources, and well-worked-out norms for their use, for handling sources of error that we *do* expect. We have

[18] To be absolutely clear, we don't mean to be saying that philosophers cannot learn anything whatsoever about the method of cases from the armchair. Surely, that would be incorrect. There has been significant discussion about the shape and structure of the method of cases. The problem is that the philosophical folk wisdom that emerges from this discussion doesn't prepare us for the many different kinds of errors that cannot be detected using only the resources that are available to philosophers from their armchairs.

[19] An anonymous referee points out to us, as indeed many philosophers have over the years, that analytic philosophers are not unaware of some version of these phenomena, perhaps most famously in Williams (1970), or more recently in Gendler and Hawthorne (2005). Our point is that whatever scraps of awareness philosophers have managed to acquire over the years about possible demographic or framing problems with intuitions, it has unfortunately not yielded really any changes whatsoever at the level of actually practiced methodology of a sort that can help us handle such influences. And as we have been stressing here, the shared disciplinary practices, not the minds of individual philosophers, are the proper target of our methodological critiques.

our norms of argument articulation, as discussed above, in no small part to help us avoid mistaken impressions of validity. Where needed, we use formal machinery such as parentheses, operators, and different forms of quotation, in order to avoid scope ambiguities, use-mention errors, and the like. When matters get particularly fraught and complicated, we can even translate the whole mess into logic and run derivations or provide models.[20]

Indeed, it's worth noting that analytic philosophical practice is not generally cavalier about errors. We don't tend to operate like big data miners, who know that tons of the specific observations in their data set will be mistaken, perhaps many of them quite substantially so, and who are thus counting on their statistical methods to help them filter the signal from the noise. Very much in contrast, our inferential and dialectical practices tend to hold our theories to a very demanding standard, something that we will return to in the next chapter. Remember that we are talking about a practice whose operating norms allow counterexamples to trump theory. Weatherson (2003) provides a nice description of this feature of analytic epistemology (and other areas of analytic philosophy):[21]

> In epistemology, particularly in the theory of knowledge, and in parts of metaphysics, particularly in the theory of causation, it is almost universally assumed that intuition trumps theory. Shope's *The Analysis of Knowledge* contains literally dozens of cases where an interesting account of knowledge was jettisoned because it clashed with intuition about a particular case. In the literature on knowledge and lotteries it is not as widely assumed that intuitions about cases are inevitably correct, but this still seems to be the working hypothesis. (p. 1)

Weatherson immediately goes on to claim that epistemologists (and other philosophers) are wrong to let counterexamples trump theory, and indeed to argue very cleverly for this claim over the course of his paper. But to the extent that his description of the inferential practices of analytic epistemology is correct, and we think he is largely on target here, this suggests that the inferential practices of analytic epistemology will be highly *error-fragile*, something that we will discuss at length over the course of the next few chapters. This means that very little threat of error is needed in order to generate the kinds of methodological concerns that

[20] We will take up the comparison between how we are trained to think about, and use, formal methods in philosophy and how we think that philosophers should be trained to think about, and use, experimental methods in Chapter 7. To be absolutely clear, we don't mean to be saying that philosophers cannot learn anything whatsoever about the method of cases from the armchair. Surely, that would be incorrect. As we discussed in Chapter 1, there has been significant discussion about the shape and structure of the method of cases. The problem, which we will discuss in Chapter 5, is that the error profile that emerges from this philosophical folk wisdom about the method of cases doesn't prepare us for the many different kinds of errors that cannot be detected using only the resources that are available to philosophers from their armchairs.

[21] See also Nado (2015) on the *epistemic demandingness* of analytic philosophy, discussed at some length in Chapter 3.

drive debates about the evidential status of the method of cases. It seems that even just a handful of bad intuitive verdicts could spoil an entire line of inquiry.

There is, thus, a clear sense in which we can say that analytic philosophical practice anticipates certain sorts of threats of error, whereas for other threats, such as the potential threats of demographic variation or instability, it is not expecting any such threats in any meaningful way. This fact about philosophical practice is absolutely essential to how we think that philosophers should think about the recent debate about philosophical methodology that has been prompted by the experimental challenge to the method of cases, and to the way that we think philosophers should think about this debate and the relationship between analytic philosophy and experimental philosophy, more generally. We will return to this over and over again in the next two chapters. For right now, it is important to see that from the point of view of our actual philosophical practices, even a very modest amount of demographic variation and instability would count as a surprising, or even shocking, level. Most importantly, so far as that first premise in the experimental challenge is concerned, there is evidence of plenty enough variation to catch philosophers napping in their armchairs and to reveal just how little we know at this time about where else further variation might yet be found.

Our diagnosis here is that Knobe tried to take a question that can only be answered by attention to the particulars of philosophical practice and construe it purely as a question about psychology. For we can observe a crucial shifting on Knobe's part in the vocabulary of these discussions, from an ineliminably philosophical one to a straightforwardly psychological one. Those who have pursued the experimental challenge have tended to define *instability* in terms like those that we see in Joachim Horvath's (2010) characterization of that challenge, what he calls its "master argument," namely, variation with "irrelevant factors." This exact phrasing can be found, for example, in the Swain et al. (2008) paper with "instability" in its title, where the authors characterize instability in precisely those kinds of terms. And Nado (2015) leads off her critical overview of the experimental challenge with a characterization in highly similar terms:

> Premise: work in experimental philosophy indicates that intuitions vary as a function of such philosophically irrelevant features as order of presentation and cultural background. Conclusion: intuitions are unsuited for their current evidential role in philosophical argumentation. We might call this the 'variation argument' against intuition. (p. 204)

It would be easy to multiply examples. One key feature of this standard construal of instability or variation is that it is *ineliminably metaphilosophical*. You cannot specify the relevant sort of variation without some idea of what sorts of factors or features are relevant to what sorts of philosophical propositions.

Yet, for Knobe, what it means for some philosophical case verdict to demonstrate instability is, simply, for it to be manipulable in ways that do not involve changing the substantive content of the scenarios. Instability is what you get when you only vary either features of the participants themselves (as with demographic variation) or in "studies on the influence of external situational factors. In such studies, researchers give all participants exactly the same case and exactly the same question, but they manipulate some factor in the external situation" (Knobe 2021, 48). Nothing about philosophical truth appears in that construal. It is entirely about what sort of experimental technique, aimed at investigating what sort of variable, is deployed in a given experimental study. To be clear, this is an obviously legitimate notion of instability, and may be just what is needed for investigating certain sorts of large-scale questions that are simultaneously scientific and philosophical about the nature of human cognition. It's just not the notion of instability in play in the arguments that Knobe has tried to declare "moot."

And once we take on the perspective of instability *a la* the experimental challenge, how various empirical findings get scored can change rather substantially. For Knobe's version of instability, content effects that are found to be cross-culturally robust get scored as instances of stability. But for the kind of instability that is central to the experimental challenge, any content effects *which arguably are philosophically irrelevant* will get scored as instances of *in*stability, and so much the worse if they are found across the board. There are several candidate effects of this sort listed in Knobe's paper, such as the influence of moral content on Gettier case knowledge attributions and the impact of physical contact in scenarios about moral dilemmas. We would expect *framing effects*, in general, to be examples of this phenomenon, as they will depend on some differences in the language or other framing elements in the contrasted scenarios, yet most typically those differences would not be ones that would be considered philosophically relevant.

Taking the point further, some demographic differences that might not seem very important for Knobe's purposes may be highly charged when we are considering their implications for philosophical practice. To take just one example, the effects of personality type that have been investigated by Adam Feltz and Edward Cokeley (2009, 2019) don't rate a mention in Knobe's 2021 paper, and while he acknowledges them in his (2023), he does so primarily to downplay their significance: "it is not as though the finding is that there is some pervasive phenomenon whereby all sorts of different philosophical case verdicts are correlated with all sorts of different individual difference measures" (p. 42). Again, it sounds like rather wild and rampant variation is what would be needed to disconfirm intuitional stability *sensu* Knobe. Yet as we noted earlier, we don't think anyone who advanced the experimental challenge ever thought that *that* much variation seemed at all likely. It's enough for the Feltz and Cokeley-type effects to be one more spear in the experimental arsenal.

The parameter of demographic variation that produces the starkest mismatch between Knobe and the experimental challengers is that of *philosopher vs. nonphilosopher*. A great many of the findings that are stable across the folk, and scored by Knobe as instances of stability to confirm his thesis, are ones where the folk's consensus verdict diverges from the received verdict in the philosophical literature. For example, Knobe's list of cross-culturally robust findings includes the irrelevance of stakes to knowledge attributions, which, if true, would be a challenging finding for much of analytic epistemology of the last few decades.[22] Philosophers' case verdicts about Gettier cases also seem to diverge in substantial and interesting ways not just from "the folk" but also from our colleagues in basically every other academic discipline (Starmans and Friedman 2020). Furthermore, a significant trend in the X-phi literature that Knobe is commenting on is that philosophers' funny thought experiments *are often kind of meh as instances or non-instances of their target concepts*. Lots of the "stability" cases are like ones where the experimental participants think that, say, Truetemp is a "kinda-sorta but maybe a bit more kinda-sorta-not a case of knowledge"—that is, exactly the sort of *inconclusiveness* cases we discussed in the previous chapter, and thus hardly results that can be of comfort to armchair practitioners.[23]

From the point of view of the philosophy of human nature, philosophers represent an incredibly small, weird (and highly "WEIRD"—that is, Western and Educated folks from Industrialized, Rich, Democratic countries) sub-sub-sample—there'd be no reason to modify anyone's views about cognitive universals based just on what esoteric whackos like us think! But, obviously, from the perspective of philosophical methodology, any "philosopher vs. non-philosopher" results will be highly fraught. We do not think that philosophers should be considered automatically wrong in any such disagreement, and we are sure there will be cases where there is good reason to let our trained philosophical judgment trump the hoi polloi. But our point here is just that instances of this particular category of demographic variation are automatically highly salient to the experimental challenge, even if they may be rightly ignored by someone with Knobe's theoretical interests.

We'll conclude this section with one last extended example. Knobe considers some very well-investigated order effects on trolley cases, particularly regarding the pattern of influences between the sidetrack case (aka "switch"; diverting the

[22] See our discussion of this sort of result in Chapter 6.

[23] Although, as noted above, we don't intend to engage with Knobe's metaphilosophical project about locating tensions in the stable differences to be found in the experimental philosophy results, we will offer this one observation: we have to be careful in interpreting inconclusive results like a 54%/46% split about *p*/not-*p*. It might be, as Knobe takes it, a manifestation of two strong and warring philosophical impulses about P, in tension with each other. But it also might be a manifestation of the folk just not having any strong views about P at all! And these do not exhaust the possibilities; for example, the difference could be the result of some philosophically shallow process. This cannot be read trivially off the distributions or histograms themselves, though we expect that strongly bimodal distributions are more likely to be indicative of tensions, whereas a big shmear centered on the Likert scale midpoint is more likely to mean that the folk just don't much know or care about P.

trolley so it veers onto another track, which will then kill one person) and the footbridge case (aka "push"; shoving the man onto the tracks, killing him but thereby stopping the trolley). In general, it appears that when someone sees the sidetrack case first, they will be much more inclined to find pulling the switch acceptable than they would when they see the scenario after first seeing the footbridge case. Interestingly, there does not seem to be any reverse effect: people just don't like killing the one to save the five in the footbridge case, regardless of whether they first see the sidetrack case. Wiegmann and Waldmann (2014) systematized and extended that literature, and proposed a model for these order effects, which we are here simplifying greatly: the sidetrack case presents a *cognitive ambiguity* that the footbridge does not. One can view the sidetrack case in a way that really foregrounds the saving of the five, and then parcels out the downstream killing of the one as a separable event, a kind of coda. Or one can view it all as one big event, with the killing included as a constituent element. Depending on which way one parses the events, and in particular whether the killing is construed as part of the action or distinct from it, one may feel differently inclined toward the action. In contrast, the footbridge case doesn't support such an ambiguity, because no matter how you parse it, you're going to have to include that killing. Thus, they hypothesize, seeing the footbridge case first leads people to be more likely to parse the sidetrack case in a manner that includes the killing, and thus to judge it as less acceptable. But since there is no corresponding ambiguity in the footbridge case, it won't experience any such order effect. And indeed, they predict that any other such trolley scenario without such a potential ambiguity of construal will similarly display no order effects, and they confirm that prediction. Knobe sums up this research as follows:

> In the specific case in which participants receive the footbridge dilemma and then receive the sidetrack dilemma, we have strong evidence for an impact of external situational factors: thinking about the first dilemma really does influence judgments about the second. What is this fact teaching us? The obvious first guess would be that it is evidence of a process that leads to some broader form of instability. Perhaps the process affects people's philosophical intuitions more generally, or perhaps just their moral intuitions, or at a very minimum, it surely affects people's intuitions about a wide variety of different trolley problems. The surprising finding coming out of research in this area is that this is not what is happening. The effect observed in this one case seems to be highly circumscribed. Not only does it not emerge in cases that are radically different, it doesn't even emerge in cases that might at first seem extraordinarily similar to the footbridge-sidetrack sequence. (2021, 66)

From the point of view of doing the scientific philosophy of human nature, that may well be right: from that perspective, and for those research interests, it might be appropriate to consider this ambiguity-based mechanism for order effects to be

"highly circumscribed," and of limited further interest. But from the point of view of the armchair method of cases, these effects should be both surprising and *very* disquieting. First, and this really needs to be underscored, even if it only afflicts a small number of cases, these cases are absolutely crucial ones in the normative ethics literature! It would already be problematic if the effects were just found on some of the odder cases, such as the one where the track loops back around on itself (Liao et al. 2012). But note that the main effect of interest in Wiegmann and Waldmann (2014) is on the classic sidetrack version of the case, and that case figures prominently in a great many of the arguments in this literature. So much of the trolley-based argumentation relies on making comparisons between the different cases, and it is problematic if we cannot, with the experimental results in, determine what to count as *the* verdict for the sidetrack case. Perhaps from Knobe's particular psychological perspective, oh, it's just this one narrow set of cases, but from the point of view of an ethicist looking to work in this topic, their whole literature will have gone off the rails.

Moreover, Knobe is somewhat underplaying how widely Wiegmann and Waldmann's mechanism for order effects may be expected to generalize outside of a subset of trolley cases. The key idea is that when cases have a particular kind of complex multi-part structure, they may enable different construals with accordingly different emphases. Now, philosophical thought experiments being the wild and chimeric creatures that they are, we ought to expect that other scenarios would manifest this sort of complex, multi-part construction. Indeed, the subclass of Gettier cases that involve switched truth-makers seem like plausible candidates for exploration here, and indeed there is some evidence of order effects on such cases (Machery et al. 2018).[24] We wouldn't be (ahem) surprised if other yet-unexamined cases did as well, whenever there's a "one hand giveth and one hand taketh away" sort of structure. Consider Alvin Plantinga's (1993) oft-used case where someone gets a lesion that causes their beliefs to be generally unreliable but *also* produces, using the same mechanism that provides lots of bad beliefs, the belief that they have a lesion. Such cases have a kind of multi-part diverging causal structure that the Wiegmann and Waldmann mechanism for order effects may well be able to get a hold of. (On reflection, it seems to us that a great many tricky reliabilism cases could potentially involve construal issues of exactly the sort highlighted by the generality problem.) Whether or not Wiegmann and Waldmann's mechanism turns out to be "highly constrained" in the specific domain of trolley problems, there is

[24] Note that the specific causal framework Wiegmann and Waldmann use for the trolley cases will not carry over to epistemology cases, typically, since ethical dilemmas tend to be about downstream consequences, whereas knowledge attributions tend to be about upstream sources. One way in which the Wiegmann and Waldmann mechanism *might* apply to the switched-truth-maker case in the Machery et al. (2018) involves an ambiguity of construal that focuses on the current state of the believer, in which they seem to check all the JTB boxes; and one that reaches further back, and registers the disconnect between the source of the justification and the distinct object that rendered the belief true.

nothing in their specification of the mechanism that would at all indicate it would not be found in many other sorts of cases in various other philosophical domains. The Machery et al. (2018) order effects seem problematic for Knobe's argument either way: either they demonstrate how the Wiegmann and Waldmann mechanism can extend in ways far outside trolley cases, or they demonstrate that there are yet other hitherto-undiscovered mechanisms in our cognition, for other sorts of order effects. This would itself not be surprising, since we already know that there are other kinds of order effects out there, such as recency/primacy effects in causal judgment (Henne et al. 2021).[25] And since causal judgments may play a role in many other sorts of philosophical domains (such as attributions of agency), there's every reason to think these could further ramify at least somewhat.

The point of these last empirical speculations is that it's entirely consistent with order effects being basically nugatory from Knobe's perspective of human intuitive cognition on the whole that they also afflict enough of the cases deployed in the method of cases to raise some very serious methodological headaches for philosophers. The problem cases don't need to comprise anywhere near a vast majority, or even a significant plurality, of the cases that philosophers use, precisely because philosophical practice is so vulnerable to them—so easily surprised by them, we might say. And the set of such problem cases is much bigger than Knobe estimates, even stipulating that he is right in his overall read of the literature, because the problem cases are delineated in terms of sensitivity to philosophically irrelevant factors, and not just the much narrower (though still decidedly non-null!) set of cases displaying purely situational effects. While we (obviously) think that psychological results are highly important to philosophical methodology, it does not follow that distinctions drawn purely in terms of psychological methodology can do the metaphilosophical work that is needed.

3 Why Metaphilosophy Should Be Meta-Methodological

We have argued that two recent attempts to respond to the experimental challenge to armchair philosophical practices from the perspective of these philosophical practices fail precisely because they are not sufficiently attentive to the particulars of those practices. But the moral of the story is broader than just this. The moral of the story is methodological, perhaps even *meta-methodological*. It is that philosophers simply cannot continue to argue about philosophical methodology

[25] We should note that this is a different kind of order effect: a content effect of the temporal order of events in the scenario, and not a context effect of the order of presentation. It is nevertheless grist for our mill since the effect is a philosophically irrelevant one, which is susceptible to manipulations that get subjects to simulate some rather than other situations in counterfactual reasoning. Moreover the authors of that study urge that their work is relevant to a wide range of philosophical issues, including moral judgment and experimental jurisprudence.

without really looking at philosophical methodology, as it is actually practiced. Having said this, we want to be careful not to over-represent or over-sell ourselves here as observers of philosophical practice. Our claims should be taken as highly empirical ones, for whose defense we have, let's face it, mostly been offering armchair observations. Perhaps we count as "participant observers," but that is hardly an innocent epistemological position. Nevertheless, we think that it is important, going forward, for philosophers engaged in metaphilosophical debates like the ones that occupy us here to look carefully at the particular twists, turns, and contours of actual, in-the-trenches philosophical practices, and we will try to model this way of doing things in what follows.

We take it, therefore, that the experimental challenge is a live one. The method of cases really is a standard and distinct part of our analytic methodology, and cannot be understood merely as an application of our other argumentative techniques yielding case verdicts as conclusions. Moreover, whether or not the total emerging picture of human philosophical cognition is one of a very high level of uniformity and stability, our practices are profoundly unprepared for even a comparatively small degree of robustness failure. So at this stage of our discussion, the experimental challenge is live—but not yet resolved either against or in favor of the armchair. In short: the game is afoot! More work needs to be done on both sides to develop the challenge to any possible point of resolution. The experimentalist needs to sharpen their normative vocabulary and determine just what kind of further premises they need for their arguments, beyond the empirical findings themselves. And the defender of the armchair needs to determine by what means they might yet parry the challenge. In the next chapter, we will take up the concepts of *reliability* and *unreliability*, and contend that despite their prominence in the literature, they turn out to be of little help in advancing the state of play, for either player.

3

The Standard Normative Framework and the Unreliability of 'Reliability'

As we discussed in Chapter 1, it is natural to think that what is at stake in recent metaphilosophical debates about the method of cases is whether the method of cases is *reliable*. After all, as Joshua Knobe (2019) writes:

> When we make frequent use of a method, it is only natural to ask whether the method is a reliable one. Suppose, for example, that we are using a method that is supposed to be 95% accurate. If this method indeed turns out to give us the wrong answer only 5% of the time, then there is no problem—the method is working exactly as it should. By contrast, if we discover that the method gives us the wrong answer 35% of the time, we would be faced with a very serious issue.
>
> The situation in armchair philosophy is no different from anywhere else. On one popular characterization, armchair philosophy makes use of a method that relies on intuitions . . . Even the most ardent defenders of this intuition-based method do not describe it as infallible. Clearly, intuition sometimes gives us the wrong answer. A question arises, however, as to whether this method has an acceptable level of reliability. If we discover that its reliability is more or less what we thought it was, this discovery would not point to a major problem. By contrast, if we find that it yields incorrect answers far more often than we thought it did, we would have reason to begin reevaluating this whole approach to philosophical research. (p. 30)

And, so, a number of philosophers have taken up the defense of armchair methods by arguing that those methods are in fact reliable. Jennifer Nagel (2012), for example, argues:

> that pre-theoretical epistemic intuitions do arise from a generally reliable natural capacity. Known to psychologists as 'folk psychology' or 'mindreading', this is our ordinary resource for ascribing states of knowledge, belief and desire. Experimentalists who challenge epistemic case intuitions do not reject all intuitive capacities—and cannot do so, on pain of collapsing into a general skepticism. At least one prominent experimentalist has explicitly identified mindreading as the kind of intuitive capacity that can be trusted, and not without reason. It is entirely plausible that this capacity is largely reliable in its deliverances, not least

Standing Up for Philosophy. Jonathan M. Weinberg and Joshua Alexander, Oxford University Press.
 DOI: 10.1093/9780191947704.003.0003

> because our intuitive mindreading generates predictions about what others will do and say, and these predictions—including predictions about the differences between thinking and knowing - are subject to feedback and correction over time. (p. 497)

And, Timothy Williamson (2007) writes in the introduction to the *Philosophy of Philosophy* that:

> [o]ne of the main themes of this book is that the common assumption of philosophical exceptionalism is false. Even the distinction between the *a priori* and the *a posteriori* turns out to obscure underlying similarities. Although there are real methodological differences between philosophy and the other sciences, as actually practiced, they are less deep than is often supposed. In particular, so-called intuitions are simply judgments (or dispositions to judgment); neither their content nor the cognitive basis on which they are made need be distinctively philosophical. In general, the methodology of much past and present philosophy consists in just the unusually systematic and unrelenting application of ways of thinking required over a vast range of non-philosophical inquiry. The philosophical applications inherit a moderate degree of reliability from the more general cognitive patterns they instantiate. Although we cannot prove, from a starting point a sufficiently radical skeptic would accept, that those ways of thinking are truth-conducive, the same holds of *all* ways of thinking, including the methods of natural science. That's the skeptic's problem, not ours. By more discriminating standards, the methodology of philosophy is not in principle problematic. (p. 3)

The basic argumentative strategy here is this: show that the method of cases is a natural extension of ordinary cognitive practices (mindreading, counterfactual reasoning, and so on) and then argue that the reliability of these ordinary practices extends out to the method of cases. And so we see Timothy Williamson observing that philosophical thought experiments "turn out to be fairly straightforward modal arguments, typically valid ones, with counterfactual conditionals and possibility claims as premises," and contending that we "assess the premises by the same . . . methods as for other claims of this sort" (2009, 433), and we see Jennifer Nagel suggesting that if "Gettier's intuitions about what Smith does or doesn't know come from our everyday mind-reading capacity for ascribing states of knowledge and belief, and if this capacity is generally reliable, then our epistemic case intuitions have some positive claim to epistemic legitimacy" (2012, 511). We also see Jonathan Ichakawa and Benjamin Jarvis (2009, 238) arguing that:

> any thought-experiment intuitions attributing concepts (whether 'conceptual analysis' intuitions or moral intuitions about trolley cases, whether a priori or not) should be accurate so long as, first, our actual faculties for recognizing

> correct or incorrect deployment of the concept are genuinely good, and second, the counterfactual cases under consideration are relevantly similar to what we might encounter in the actual world. Hence, we have some reason to think that any thought-experiment intuition should allow for knowledge so long as we are sufficiently good at deploying the concepts constituting the propositional content of the intuition in actual cases and so long as the thought experiment isn't so very far-fetched.

While this strategy for defending the armchair may seem straightforward enough, we contend that such broad reliability claims provide an inadequate frame for debating the relevant methodological questions—both because of differences in the reliability of case verdicts from ordinary settings and purposes to those of philosophical inquiry, and also because such an on-the-whole reliability claim would, by itself, be inadequate to establish armchair philosophy's methodological self-sufficiency. But before we argue those points, it will be helpful to say a little bit about what philosophers often mean when they say that sources of information and other belief-forming processes and practices are reliable. In ordinary use, *being reliable* is basically synonymous with *being trustworthy*, and applies not only to epistemic contexts, where we talk about things like reliable sources of information, but to other contexts, as well, where it's common to talk about someone's being a reliable friend or a reliably good guest at a dinner party. Someone who is reliable in this ordinary sense is someone who can be expected to "hit the mark" some suitably large percentage of the time, where what this means will be relative to domains and purposes. Reliably good drivers had better not get into an accident more than a vanishingly small percentage of the times that they get behind the wheel, for example, whereas a baseball player who hits safely only 40% of the time is a fabulously reliable batter. Philosophers usually have something a bit more specific in mind when they talk about reliability, although what they have in mind has the same kind of domain- and purposes-relative normativity built in. When philosophers, and especially epistemologists, talk about reliability what they have in mind is a simple measure of how likely sources of information and other belief-forming processes and practices are to produce true beliefs, where reliability is measured on a somewhat standard scale with a floor somewhere significantly above fifty percent and a threshold probably not greater than ninety-five percent and certainly not as high as one hundred percent, and where being sufficiently reliable is again relative to domains and purposes.[1] This kind of reliability is, arguably, what Goldman (1979, 10) has in mind when he says, "(on first approximation)

[1] Importantly, this philosophical use of "reliable" diverges from the use of "reliable" common in other academic disciplines, in which a reliable test is one that will keep giving highly similar results when applied again and again to the same sample, even if it is in fact objectively inaccurate. Philosophers would be more likely to call such a test "self-consistent." Our thanks to Joan Weiner for her insistence on this point.

reliability consists in the tendency of a process to produce beliefs that are true rather than false," and it has played a significant role not only in epistemology, where philosophers have used this notion of reliability to build theories of knowledge and justification, but in many other debates in philosophy. The "rationality wars," for example, may well be fought over whether or not our ordinary inferential capacities are sufficiently accurate (see Samuels et al. 2002). And even debates about eliminativism about propositional attitudes can perhaps be understood in terms of whether our descriptive and predictive folk-psychological practices are more or less accurate or wholesale mistaken. It is important to note that reliability is a fairly coarse-grained concept, but these are philosophical debates that can operate at a fairly coarse-grain level. Even champions of propositional attitudes, such as Fodor, would be happy to grant that we can often be wrong in our attributions of beliefs and desires, arguing that these mistakes need to be viewed against a background of a pretty high level of overall success.

1 Why Reliability Is Not Enough

With this in mind, let's suppose that something like the view that Nagel and the others advance is right, and that the method of cases is generally reliable in the sense that these methods are simply extensions of reliable cognitive processes and practices.[2] And let's call this the *general reliability thesis*. Importantly, the general reliability thesis does not entail any more specific reliability thesis, in the following sense: different applications of some epistemic capacity to different materials, or in different contexts, will not be uniform in their reliability. Reading a highway billboard on a clear sunny day, and calling balls and strikes for 98 mph pitches right near the corner of the strike zone during a drizzly night game are both applications of sense perception, and yet the umpire's capacity with the former can nonetheless be expected to be much more reliable than the latter. And that leads to the trouble for the general reliability thesis as a means of deflecting the experimental challenge and establishing the intended result of the methodological autonomy of armchair philosophy. Put another way, the trouble with the general reliability thesis is that it turns out that what is good enough for ordinary epistemic purposes is not always good enough for other specialized purposes, including not just professional sports officiating but also those of academic research methodology. The reasons why are quite simple. Sticking for the moment to the realm of evaluations of reliability and unreliability, when it comes to methodological questions, how reliable

[2] We will not press the point, but we would note that we cannot stipulate that they are *all* correct, as their views do not seem to all agree as to which part of our ordinary cognition the method of cases is an extension of (i.e., folk psychology, counterfactual judgment, conceptual applications).

a belief-forming practice has to be in order to be reliable *enough* is relative to domains and purposes. Notice the important caveat in the Ichikawa and Jarvis passage that we quoted above, namely, that the kinds of "cases under consideration are relevantly similar to what we might encounter in the actual world." And there are reasons to think that philosophical practice is sufficiently *dissimilar* to the kinds of ordinary cognitive practices that Nagel and others appeal to when defending the general reliability thesis. To demonstrate this, we will follow Nagel's lead and focus on three ways in which how case verdicts get used in epistemological research diverges from our ordinary practices with epistemic verdicts; we expect that what we say here about the ways in which epistemology differs from ordinary epistemic practices will generalize well to the relationship between armchair philosophy and ordinary cognition, more generally.

The first way in which epistemology diverges in important ways from ordinary epistemic practices is that epistemologists are frequently interested in cases that involve extensive details that we suspect are not *ecologically valid*. Very often, these cases include highly specific information about an agent's mental states and processes, together with some story about how these states and processes connect with other features of the fictional vignette, where this information may not be available to the agent—and, most importantly, is of a sort almost never available in the real world. Gettier cases are good examples of this, and it is important to keep this feature of these cases in mind regardless of whether or not real-world analogues can be constructed for many hypothetical thought-experiments (Williamson 2007). Our ordinary capacity to think about epistemological issues has likely been shaped to evaluate situations where we typically have rather sparser, noisier access to what might be going on in someone's head, and this mismatch between epistemological thought experiments and the proper domain of our ordinary epistemic capacities puts serious pressure on the suggestion that the general reliability of our ordinary epistemic practices extends to how epistemologists use the method of cases.

Of course, a defender of the armchair might well suggest that our knowledge verdicts on such cases will be even *more* reliable than our verdicts in more typical cases, because we have all this extra information stipulated into the vignette to draw upon about the agent's thought processes and so on. Such a claim may seem obvious, but, in fact, extra information can often be a distraction or, even worse, a poison pill for various unconscious judgment processes; see, for example, discussions of "less is more" effects in the ecological rationality literature such as Gigerenzer and Brighton (2009). Nonetheless, it does indeed seem to us to be a plausible hypothesis, well worth investigating. But at this point such a claim would be just a mere conjecture, and about an empirical matter which clearly falls outside the investigatory scope of armchair methods. Our main point stands: there is no easy, armchair-available inference from general reliability in ordinary sorts

of cases to a similar degree of reliability in the sorts of cases philosophers like to traffic in.[3]

The second relevant way in which epistemology diverges from our ordinary epistemic practices is that epistemologists want to do something with our knowledge verdicts that ordinary folks are not typically interested in doing, namely, using them to argue for some preferred epistemological theory or another. And for this reason, they sample esoteric, complicated, and far-flung sorts of cases to a much greater extent than our ordinary epistemic practices do.[4] As we noted briefly in our discussion of the method of cases in Chapter 1, it is widely acknowledged that such factors may impair the reliability of case verdicts. Although Gettier cases are perhaps the most famous hypothetical cases in the history of epistemology, they are certainly not the only ones. The history of epistemological research is full of important cases, and our verdicts about such cases are often meant to place substantial constraints on our epistemological theorizing: to comply with their verdicts or face a nontrivial burden of otherwise accommodating them when we wish to dissent from them. Some of these cases are fairly ordinary, such as cases where we are asked whether someone knows in advance that she has lost a lottery simply because she knows that the odds of having won are miniscule. But very often they are much more esoteric, such as cases involving evil demons (old and new), or clairvoyants, or bizarre brain lesions with weird effects not to be found even in the works of Oliver Sacks. There's a good methodological reason that such high-flying and funky cases are both common and important in analytic epistemology, for we are often trying to get some evidential traction in the slippery effort of rationally preferring one from a set of competitor epistemological theories, all very good and all rivalrous with each other. And so it doesn't even matter that much whether epistemologists sometimes, or even very often, deploy more ordinary cases, so long as they also rely crucially on more esoteric cases—when they do so, the general reliability thesis can offer no reassurance to them.

Furthermore, the very factors that make verdicts about unusual cases methodologically crucial also lead us to positively expect them generally to be more error-prone.[5] They will be more susceptible to subtle effects of context, for example, because they will often involve splitting apart distinct features that commonly go together in knowledge attribution and using them against one another, such as the baseline accuracy of a piece of cognition and the availability of at least some considerations that speak in favor of that accuracy. Different contexts may cue up different weightings of such features in our unconscious categorizing systems and, thus, have an increased potential to produce different attributions. As for group

[3] Our thanks to Jennifer Nado for this point.

[4] See Alexander and Weinberg (2007); Weinberg et al. (2012); and this whole section borrows heavily in particular from Alexander and Weinberg (2014). See also Machery (2017) for some parallel arguments about the nature of philosophers' preferred cases.

[5] Machery (2017) calls these *disturbing characteristics*.

differences in cognition, that stands as a vibrant and hotly debated topic in psychology these days, and is, we think, a rather more open question than some epistemologists take it to be. But we do not need to take sides in that ongoing debate here. For our purposes, it is enough to note that in order to raise worries about the relevant philosophical practices, one need not go as far as Nisbett, Peng, Choi, and Norenzayan (2001) or Henrich, Heine, and Norenzayan (2010), and claim strong and pervasive differences between WEIRD and the rest of the world. For example, although he is a prominent critic of those authors and their claims of deep group differences in cognition, even Mercier (2011) suggests that we should still expect to find fine-grained cultural differences in knowledge attributions, even against a widely shared background of cognitive convergence. We suggest that these more subtle differences in folk epistemologies will be more likely to manifest in the unusual and marginal sorts of cases that are popular with epistemologists, than in ones that are evaluated and discussed in some frequency in civilian life. It is thus consistent with high baseline accuracy for epistemic case verdicts in general that the sorts of case verdicts that are important in analytic epistemology may suffer more from unwanted and unanticipated sensitivities.

The preceding two arguments concerned ways in which the verdicts appealed to in the method of cases may demonstrate a baseline degree of reliability inferior to that of the capacities that they are an extension of, according to the general reliability thesis. But the third way in which epistemology diverges in important ways from ordinary epistemic practices has to do with not just the kinds of *cases* that we frequently look to deliver verdicts about, but also the kinds of *inferences* for which those verdicts are meant to serve as premises. Nagel concludes with the claim that epistemic case verdicts are "reliable enough." But one question that must be asked here is, reliable enough *for what purposes*? Even if (*pace* the above considerations) the armchair method of cases displayed the same baseline reliability as the rest of our ordinary cognition, that may not be a degree of reliability appropriate to the uses to which philosophers put it. For most of the purposes involved in everyday cognition, we are inclined to agree with Nagel that epistemic case verdicts are probably trustworthy enough to guide us in our quotidian transactions with the world. But our philosophical purposes are far more exacting, and one clear way to see this is in the kinds of inferential uses to which epistemic case verdicts are put in epistemology, but not in ordinary life. It is not an exaggeration to think that every epistemological theory that is based at all on our epistemic case verdicts is based on sophisticated inferences driven substantially by those case verdicts—no one thinks that they can just read the correct epistemological theory directly off of the cases, non-inferentially. As such, to the extent that we are interested in questions about baseline accuracy here, it seems that we should not primarily care about the baseline accuracy of the epistemic case verdicts themselves but rather about the conclusions that will be inferred, based on those case verdicts. We do not just want our theories to be based on premises that are mostly true; we even more want our

theories themselves to be mostly true. What is at issue is not the obvious point that inferences should be conditionally reliable, but whether conditional reliability is enough, and there is reason to worry that it is not, and that some additional dimension of epistemic evaluation is needed. The conditional reliability of a rule of inference is a matter of how likely a conclusion is to be true *assuming all its premises are true* (Goldman 1979). But different rules can do better or worse at handling situations where errors have started to creep into the set of premises.

To see why, consider two inference rules that are equally reliable on all true inputs but whose reliability diverges sharply as soon as the quality of the inputs begins to degrade: one inference rule that takes ten propositions as inputs and outputs their ten-way conjunction and another inference rule that takes the same ten propositions as inputs and outputs their ten-way disjunction. These two rules are maximally conditionally reliable; when all of the inputs are true, both will produce true outputs all of the time. Yet when the quality of the inputs degrades, so does their reliability—but not in the same way, and that is where this other dimension of evaluation can be seen. For the ten-way conjunction rule becomes maximally conditionally unreliable as soon as any of its inputs deviate from the truth, whereas the ten-way disjunction remains maximally conditionally reliable so long as at least one of its inputs is true. This dimension of the evaluation of rules of inference closely resembles a related issue in the evaluation of models, in which we evaluate them in terms of robustness, or how well they withstand alterations to their basic assumptions and parameter settings. So, let us call this dimension of inference evaluation *error robustness*. Here we want to suggest that part of determining whether or not the method of cases is trustworthy involves determining how error-robust the epistemological inferential practices are that take its case verdicts as premises: the more *error-fragile* one's inference rules are, the less tolerant of unwanted sensitivities one can afford to be in one's premises.[6]

And there is good reason to worry that the inferential practices of analytic epistemology are highly error-fragile. Jennifer Nado (2015) talks about this in terms of the *epistemic demandingness* of different sorts of inquiry, and argues that philosophical inquiry, which so often is framed in terms of exceptionless universals, is enormously demanding. That is, many modes of philosophical inference, including philosophical analysis, require a much higher degree of reliability than other modes of cognition, especially many that operate just fine for the purposes of our ordinary lives. She illustrates this with the following useful example (2015, 213–214):

[6] In Chapter 2, we discussed error-fragility in terms of how philosophers use counterexamples, focusing on Weatherson's (2003) influential discussion of how philosophers often let counterexamples trump theory. Error-fragility will be a central part of our discussion of methodological rationality in the next chapter, and as we've just noted connects back to our discussion of *robust case verdicts* in Chapter 1.

> Consider a group of 10 objects, *a, b, c . . . j*, and two properties, *F* and *G*. Now consider a subject who possesses a 'folk theory' devoted solely to those objects and their properties, on the basis of which the subject makes judgments regarding the applicability of *F* and *G* to the objects in the group. Suppose that, by means of this folk theory, our subject produces the judgments *Fa, Fb, Fc . . . Fj*, and the judgments *Ga, Gb, Gc . . . Gj*. Finally, suppose that in actuality, ~*Fa* and ~*Gb*—all other judgments are correct. Out of 20 judgments, the subject has made 18 correctly—she is, then, a reasonably reliable judger of *F*-hood and of *G*-hood on the cases to which her folk theory applies. We would likely say that it is epistemically permissible for the subject to rely on such judgments in normal contexts.
>
> Suppose, however, that our subject is a philosopher; further, suppose her to be concerned with the nature of *F*-hood and of *G*-hood. Our subject might then come to hold certain theoretical claims about the nature of *F*-hood and *G*-hood on the basis of those initial classificatory judgments. She might, for instance, infer that everything (in the toy universe of 10 objects) is *F*, that everything is *G*, and that if something is *F* then it is *G*. She would be wrong on all counts. The example is simple, but it shows that a certain principle—that the general reliability of one's classificatory judgments directly entails the general success of one's theory-building—is clearly false. Generating an accurate theory is highly epistemically demanding; an otherwise respectable source of evidence may not suffice.

She proceeds to argue in her (2016b) that this is one of the reasons that philosophical inquiry has *higher standards of reliability* than ordinary epistemic practices. She makes this point in order to defend our ordinary epistemic practices from the kind of skepticism that would result from applying these higher standards to those practices, but for our purposes here, what it is important to see is that this also means that whatever features of our epistemic verdicts make them reliable enough for the purposes of those practices, they aren't necessarily going to make those verdicts reliable enough for the purposes of philosophical practice.

While we wholeheartedly agree with Nado's argument for the methodological inadequacy of general reliability, we nonetheless dissent from her proposed replacement with simply a supra-normal standard of reliability. In the end, we think that the concept of reliability, even dramatically strengthened, simply cannot do the needed methodological work. Put another way, even a much higher level of reliability may not be enough to establish the trustworthiness of a source of evidence. To see this, just consider the above thought experiment, but then ramp up the numbers so that the Nadonian judge gets 98 out of 100 trials correct, with the incorrect *Fa* and *Gb* verdicts. Or, say, 498 out of 500. Nado's argument would still go through just as well. Maybe there are some stratospherically high-enough numbers here, where the proper move would be to say, as it were, 50,000,000 *F*-ness fans can't be wrong, and legitimately refuse to count that one ~*Fa* as a problem case. If so, then perhaps we could follow Nado on the condition that we insist on a *very*

high standard of reliability—much higher, surely, than we could currently affirm for armchair case verdicts, given the state of the psychological evidence.

While this line of argument is sufficient to problematize Nagel's response to the experimental challenge, we would nonetheless go further. For maybe even *very* high reliability is not enough, and indeed we have a deeper worry that proportions of right verdicts just isn't a fully adequate way to think about this issue. Consider Gettier cases, and the current state of play in epistemological theorizing. Now suppose, just *arguendo*, that the situations we classify as Gettier cases really are, in fact, instances of knowledge—but that we are superbly reliable about every other kind of case. On the one hand, Gettier cases are a truly miniscule proportion of the total set of situations that we could be called upon to evaluate as knowledge or not. If we're wrong about those cases, and *just* those cases, then the baseline reliability of our epistemic case verdicts would surely be very high indeed. On the other hand, because of the theoretical centrality of such cases in the last half century of epistemology, on this hypothetical *nearly every single theory of knowledge currently on offer would be wrong*.[7] In terms of their proportion of the total verdict set, we'd only be wrong on a rounding error's worth of possible cases—but even so, the error-fragility of our methods mean we'd still be facing an utter methodological disaster. What we seem to need is not a standard of *even-more-astronomically-reliable* reliability, but something cut from an altogether different conceptual cloth. We will discuss this issue in greater depth in the next chapter; in short, what's needed is not so much tightening up the reliability standards on the front end, but making sure on the back end that we can hunt down and rectify any errors that might have been made. But the key point here is that not only is Williamson's "moderate degree of reliability" inadequate in and of itself to secure the armchair's methodological autonomy, so too is Nado's significantly heightened degree.

Let's take stock. We agree with Nagel that the general reliability thesis is highly likely to be true, both about our specifically epistemic capacities but also about our verdictive ones at large. Nonetheless, we disagree sharply about how much work such a thesis can do in debates about philosophical methodology. Differences between the conditions and demands of philosophical inquiry lead, for good reasons, to substantial differences in the kinds of cases that get considered for verdicts, and the kinds of inferences that are meant to be served by those verdicts. As a result, there is no reason to think that the verdicts as used in the armchair method of cases are quite as reliable as those used in our ordinary practices. Moreover, even if these verdicts were on average as reliable as ordinary cognition on the whole, we have reason to suspect that the degree of fallibility in such verdicts is one that poses a great methodological threat to the error-fragile inferences to which they are put, given the strenuous demands of our theoretical ambitions. Either one of these lines

[7] At the same time, Kaplan (1985), Sartwell (1992), and Hetherington (2012) would need to be awarded some special sort of medals for epistemological valor.

of argument would be problematic for a reliability-based defense of the armchair; together, they are fatal. General reliability is of course a very fine thing, as far as it goes. It just turns out not to go very far, in debating methodology.

2 Why Unreliability Is Too Much

While it is not enough for the armchair defender to argue that their methods are simply extensions of ordinary, and generally reliable, cognitive practices, it might still be tempting for the experimental challenger to think that it would be a good strategy to try to show that these methods are *unreliable*. After all, it's natural to look at evidence for problematic heterogeneity and instability in people's proffered case verdicts, and take it all to count substantially against the reliability of the method of cases, and infer from there to a rejection of the method of cases.

Edouard Machery (2017) presents the most careful and sustained version of this unreliability argument, and argues that since the method of cases plays a seemingly ineliminable role in helping philosophers answer the kinds of questions that interest them, accepting this methodological prescription means abandoning not only the method of cases, but also many of the questions that philosophers have come to associate with philosophy itself. It is unsurprising that Machery's "bounded philosophy" has philosophers worried, and a number of thoughtful responses have already been raised about some of the main argumentative moves that he makes in support of his claim that the method of cases is unreliable. Janet Levin (2019), for example, has argued that Machery moves too quickly from his evidential basis, that is, the kind of work already done by experimental philosophers, to his inductive conclusion that armchair philosophy is unreliable, especially because some of the work already done by experimental philosophers seems to suggest that heterogeneity and instability occur to varying degrees across different kinds of philosophical discussions; to take just one example, work done by experimental philosophers suggests that demographic variation is less likely about epistemological cases than ethical ones. And Michael Strevens (2019) has argued that, while Machery gives us reason to think that heterogeneity and instability are systemic, he has not done enough to show that verdicts about philosophical cases are *incurably* unreliable. Strevens approaches the issue through the lens of different theories of concepts and argues in different places that our verdicts about philosophical cases cannot be incurably unreliable either because the problematic kinds of heterogeneity and instability associated with philosophical cases should be compartmentalizable, or otherwise correctable with careful attention, better concepts, or better 'inferential chops', or because stable, widespread heterogeneity and instability would be evidence that there simply is no fact of the matter.

These kinds of responses share a key assumption with Machery: that the notions of reliability and unreliability are the terms on which the debate about

philosophical methodology should be staged. And we want to reject the terms of engagement. But before we do, we need a preliminary canvass of just what flavor of unreliability Machery has in mind; that is, what ratio of correct answers to incorrect constitutes the threshold between epistemic respectability and evidential ruin? He doesn't say, exactly, but we can get a sense for the general contours of his view from some of the things that he says at various points in the book about reliability and unreliability: he counts 77.5% correct as straightforwardly reliable (104), whereas 55% is unreliable, at "barely better than chance" (105). Somewhere in between, 62.5% is "sufficiently low to question . . . reliability" (108). (Knobe seems to be on a similar wavelength here, in the quote that opens this chapter.) In effect, Machery attempts to avoid staking himself to any sharp boundary between reliability and unreliability by focusing instead on a range of what might be called *moderate unreliability*: when a source is reliable enough that one ought not simply dismiss its evidential weight, but not so reliable that one can just trust it uncritically. While we don't think that this will ultimately work, for reasons that we will talk about in a moment, it is worth mentioning here that this seems to be a move that he *has* to make—because moderate unreliability is very likely the only kind of unreliability that would have any chance of working for him. Set the threshold too low, say right at 50%, and it really will not be at all possible to demonstrate that the method of cases is *that* unreliable using anything like the kind of work that has been done by experimental philosophers. And set the threshold too high, say at 95%, and almost no one would be willing to grant that such minimal unreliability would be grounds for deep concern.[8]

With all of this in mind, we think that focusing on unreliability (even moderate unreliability) just doesn't work and that the problem is baked right into the very notion of unreliability, at least as that concept is used by philosophers in these debates. In order to support the claim that the method of cases is unreliable, even moderately so, Machery needs to be able to show that there are no more than n correct judgments out of the total of m cases in the target class, such that n/m is not much greater than 0.5.[9] And there are at least two problems with trying to establish this kind of claim empirically. First, there is the *numerator problem*. Being able to calculate the ratio of accurate judgments to inaccurate ones would mean already having some way of reckoning that n, which requires a separate way of distinguishing true judgments from false ones, something we sadly don't already have. After all, if we did have some independent way of doing this conclusively, there would be no need for such fuss about the method of cases. Second, there

[8] It turns out to be deeply hard to find a degree of epistemic flaw that avoids this problem; see the discussion of the Scylla-Charybdis difficulties with the "impeachment schema" in the next chapter.

[9] See below on how these notions need to be understood modally. That is, n and m almost certainly ought not be calculated from the set of merely actually occurring thought experiments, but something more like the set of potential thought experiments that cross some threshold of likeliness of occurring in the method of cases as currently practiced.

is the *denominator problem*. Empirical work on the method of cases can perhaps show that this or that case is problematic in this or that way, but without some way of determining how big the reference class *m* is, we are simply not in the position to calculate the ratio at the heart of Machery's unreliability argument. Machery is clearly aware of these problems since we see distributed throughout his argument various attempts to bypass them; instead of trying to find a way, *per impossibile*, to determine these values, he offers clever proposals for doing the same work, but with different materials. Yet the numerator and denominator problems prove to be not so easily circumvented.

The Numerator Problem

Let's start with the numerator problem: to whatever extent we have an established method already ready at hand for distinguishing correct verdicts about philosophical cases from incorrect ones, they clearly don't extend nearly far enough to cover more than a fragment of the range of the method of cases and so there is simply no available direct way to calculate *n*. Machery thinks that we can avoid this problem by focusing on *effect sizes*: large enough effects mean that people in one group or context will give such different responses from those in other groups or contexts that at most only about half of them could be right. And that would entail unreliability even in conditions where we don't know what the right answer happens to be. Here's a key passage from chapter 3 of his book, with "*V*" standing in for some demographic or contextual variable of interest and "*J*" for a judgment about a case:

> When does the influence of *V* on *J* mean that *J* is unreliable? If *V* has a large influence on *J* and if *J* is dichotomous (e.g., Is it permissible to push the large man in the situation described by the footbridge case? Yes or No?), then judgments are unreliable. Suppose that 80 percent of liberals correctly assert that global warming is real for only 30 percent of conservatives. If people are equally likely to be liberal and conservative, then 55 percent of people are likely to get it right, barely better than chance. So, the judgments elicited by a given case are unreliable provided that they are influenced by at least a demographic variable or a presentation variable and provided that this influence is large. (p. 105)

This is a clever move, for sure. But it ultimately will not work, especially when we think about the kind of work that experimental philosophers have done and the kinds of demographic characteristics that they have focused on. There are at least two reasons. First, the move sketched in the passage above works only when relevant demographic groups are at least roughly equally represented in philosophy. See that crucial stipulation in the quoted passage that *people are equally likely to*

be liberal and conservative. Machery needs this kind of stipulation.[10] To see why this stipulation is simultaneously crucial and fatal, it is important to see that it helps Machery handle a different kind of denominator problem than the one that we will be pressing below; namely, that since reliability is a modal notion, the set of possible cases in the reference class could be intractably open-ended, perhaps even transfinite, rendering the requisite ratio potentially incalculable. Machery avoids this kind of denominator problem by relativizing reliability to some particular environment so as to keep the cardinalities at play manageable. But once he introduces an element of relativization to some particular environment, the kind of stipulation he's making here becomes necessary. In order to evaluate reliability in some particular environment we need to consider how specific characteristics are likely to be distributed within the relevant environment. But here's the rub: in the context of evaluating the reliability of the method of cases for philosophical inquiry, the relevant environment has got to be the environment of philosophical practice; yet philosophical practice today certainly seems disposed, in an unfortunately robust way, to vastly oversample from some specific demographic groups and not others. As such, the stipulation of an even distribution of the variable's values across the population is one he needs, but he cannot have.[11]

The second reason why Machery's focus on effect sizes won't work emerges when we shift our focus from demographic effects to context effects. Again, Machery would need it to be the case that the different causally influential contexts tend to be at least roughly equally likely to be encountered by working philosophers in the environment of philosophical inquiry. And this is both unlikely on its face and highly non-trivial to investigate. Suppose one finds (as Machery et al. 2018 has) a significant framing effect on Gettier cases. *Maybe* those frames are equally likely to be encountered, but surely there is no easy way to figure that out

[10] In addition to the reason that we will focus on in what follows, there's another—and perhaps even more obvious—reason why Machery needs to make this stipulation. To see why, consider the hypothetical case he describes in the quoted passage, but where the population is nine hundred liberals and one hundred conservatives. In that case, we would have 720 liberals correctly asserting that global warming is real and thirty conservatives correctly asserting that global warming is real, which means that 75% of the population are getting it right. Not so unreliable after all! (And notice how *deceptively* easy it is to calculate reliability when you can just stipulate the size of the population and how many of them are getting things right.)

[11] For some variables, Machery could perhaps stipulate that the modal space could be legitimately *idealized* away from current demographic imbalances; most obviously, we could try to stipulate that men and women be equally represented in the suitably idealized space of philosophical practice, even though as a matter of current fact that is far from the case. Yet it is hard to see how to generalize from dichotomous categories that are roughly proportional in the population to more complicated cases. According to some estimates, roughly ⅓ of the population are introverts, and ⅔ are extroverts; supposing that's correct—and we know that there are philosophically meaningful variations in intuitions along this variable—do we consider introverts at ⅓ of the space, or as equally likely as extroverts? Of course, it gets more complicated for non-dichotomous variables. In the idealized modal space, what fraction of the possible practitioners should be white, and what fraction for other races? Ditto for native English speakers versus native speakers of the many, many other linguistic groups who may currently be underrepresented in analytic philosophy.

from the armchair, which is all Machery is going on here. And, while investigating this scientifically is at least possible, it would involve a lot of a very different kind of empirical work than what has generally been done by experimental philosophers, combing carefully through the published oeuvre of Gettier cases and evaluating the frequency with which the different frames are used. That would be hard enough; yet the tendencies toward different comparative incident rates of other sorts of contextual influences, such as environmental ones, are very likely to be downright empirically intractable.

What's more, for at least some kinds of context effects that have been studied by experimental philosophers, we can make a plausible armchair estimate that these kinds of contexts are ones that philosophers will find themselves occupying so rarely that even a huge effect size could not amount to overall unreliability. For example, there is some evidence that conducting a thought experiment written in a hard to read font like *Mistral* can influence verdicts about free will cases (Weinberg, Alexander, Gonnerman, and Reuter 2012). We think this effect indicates a worrisome broader sensitivity to metacognitive fluency cues, and that's why it's relevant to the critique of the armchair. Nonetheless, this specific "weird font" context is one that philosophers are not at all likely to encounter with much frequency in the environment of philosophical inquiry—outside of reading the occasional X-phi study, that is. To make just one rather simple analogy, being struck by lightning would likely have a rather large influence on both the case verdict reliability and short-term mortality of philosophers, but (thankfully) it has no propensity to happen all that often.

Another problem that we have here is that Machery's argument requires both that the conditions are roughly equally likely and that the effect across the conditions is large; however, many of the relevant effects that have been reported by experimental philosophers have not had a large effect size.[12] Machery says that he is not concerned about small effects because he thinks that they can "add to one another" (2017, 108). Get enough small effects, and it is just as much a threat to reliability as one large one, or so he argues. The problem, though, is that if small effects can be additive, then they can also be subtractive. If there are a number of independent small effects operating on a dichotomous judgment, then Machery is right that sometimes they will all push in the same direction, giving a larger total effect. But much more often, we would expect some will pull against others, perhaps not canceling each other out completely, but nevertheless leaving a less dramatic effect overall.

Here's a toy example to illustrate the point. Suppose that there is a case in epistemology where most people, say 70%, are willing to attribute knowledge, but that there are three factors that can influence how likely people are to attribute

[12] For an interesting recent discussion, see Demaree-Cotton (2016).

knowledge in this case. And let's suppose that each of these factors is independent of the others and each is just as likely to influence people's willingness to attribute knowledge in this case. We will also follow Machery's stipulation that each factor is equally likely to occur in either setting. To make things slightly easier, let's call these three factors F1, F2, and F3 and add a + or - to indicate when they increase or decrease the likelihood that people will attribute knowledge in this case. So, for example, in cases where F1+, F2+, and F3+, we'd see a 30% increase in people's willingness to attribute knowledge, and in cases where F1-, F2-, and F3-, we'd see a 30% decrease in people's willingness to attribute knowledge. Both of these extreme combinations would be philosophically significant; one combination making it certain that people would attribute knowledge in the case, and the other making it unlikely that people would attribute knowledge in the case. Let's represent these possible combinations and the resulting rate of attribution/non-attribution in the following way:

<F1+, F2+, F3+> = 100:0
<F1-, F2-, F3-> = 40:60

Machery wants to use these kinds of possible combinations to show that factors that have small effects can, nonetheless, line up in the right way to produce meaningful differences. The trouble is that these two possible combinations are not very likely. After all, given three independent factors:

p(F1+ & F2+ & F3+) = 1/8 = p(F1- & F2- & F3-)

What's much more likely (in fact, three quarters of the time) is that two factors will combine to increase people's willingness to attribute knowledge in the case while the other one will decrease people's willingness to attribute knowledge, or vice versa, which would result in a net increase or decrease of only 10%. People would be slightly more likely or slightly less likely to attribute knowledge, but not in any way that is of any use in arguing for an unreliability thesis. After all, in this simplified case, only ⅛ of people will offer a "not knowledge" verdict. This is just a toy case, of course, but we hope that it illustrates the general problem, for Machery, with hoping that small effects will aggregate into large ones. Although in some sense they can do so, the circumstances in which they will do so will typically be much rarer than the circumstances in which they flatten each other out.

There's something else worth mentioning about this toy example, which connects back to the passage we quoted above; namely, that despite the fact that Machery frames his argument in terms of effect size, it looks like effect size per se is not all that important. What seems to be much more important for Machery's argument is the distribution of judgments and to what degree on average they fall above or below some break-even point. Consider, for example, a case where all

liberals agree that A, but only 60% of conservatives do. (Perhaps let A = Joe Biden fairly won the 2020 general election.) That's a sizable effect; yet the population on the whole (if we stipulate an even distribution) agree that A. Conversely, if the mean is right around break-even, then any changes brought about by a factor with even a modest effect size could still be enough to make the population unreliable. So, as we noted in our discussion of inconclusiveness in Chapter 1, perhaps the crucial finding in many experimental philosophy studies is not that there are large effects but that people's judgments about these cases average out pretty close to the midpoint.

While we are focusing on what Machery says about the inductive basis of his case against the reliability of the method of cases, we also want to mention a worry that we have about some of the things that Machery says about *partitioning*. Here Machery considers the possibility that unreliability might be quarantined sufficiently if philosophers could focus only on certain kinds of philosophical case verdicts. For example, it has been suggested that if we attend only to expert philosophical case verdicts, perhaps these sorts of errors will not arise nearly so much. Machery thinks that the problem with this strategy for minimizing worries about unreliability is that we know so little about what kinds of partitions make a difference. We heartily agree with Machery about that, and discuss this more in Chapter 5. The problem is that we don't think that he can make much argumentative appeal to this state of evidential impoverishment regarding various partitions, for reasons that we will address in more detail in Chapter 7. Machery needs to be in a position to claim to have established that the method of cases is unreliable, a claim he makes at many points in his book. The best that he could get from this kind of no-evidence-yet strategy would be something like '*for all we know*, the method of cases is unreliable', which is obviously much weaker than the needed unreliability claim itself. Machery seems sensitive to this distinction. He writes, for example, "as far as we have information, reliability is invariant under partitioning . . . This could happen for one of two reasons: first, reliability is invariant under partitioning; second, while reliability varies under some partitioning, we have no information about it" (p. 99). The problem is that Machery, in order to actually secure his key unreliability claim, very much needs it to be the first reason, and we just don't know at all yet whether that's true, which is a big part of the reason why we think that there's still really interesting work for experimental philosophers to do.

Now, we agree with Machery that we have good evidence that some ways of partitioning don't make a difference, especially with respect to the so-called expertise defense and the kinds of problematic sensitivities that experimental philosophers have been concerned with.[13] But there's good reason to think that there are probably a number of yet-unrevealed moderating factors out there; for example,

[13] We will talk more about the expertise defense in Chapter 5. Nick Byrd (2022) makes a similar argument.

appropriate kinds of reflection and motivation probably mitigate *some* problematic kinds of sensitivities, exacerbate others, and leave others, perhaps most, entirely untouched. The point is that not only do we not know what those are yet, we are unlikely to be able to determine this from the armchair. And that state of play should work just fine on behalf of proponents of the experimental challenge, since armchair-unresolved—and indeed armchair-*unresolvable*—worries about error are our stock in trade. But all in all, this means that Machery doesn't yet have the right to his much-needed empirical premise. One thing that makes this problem particularly, well, problematic for Machery is this: to undermine his argument as he runs it, any such yet-undiscovered moderating factors wouldn't even need to come close to making the problematic effects go away in order to defeat his unreliability claim; they just need to reduce the effect size from large to moderate. After all, Machery's unreliability claim is much more ambitious than merely the claim that things are just somewhat noisier than we might have liked them to be. And so he really needs to be able to rule out quite a lot that, it seems to us, is not yet ruled out by the empirical work that has been done, at least not yet.

The Denominator Problem

So much for the numerator problem. What, then, about the denominator problem, that is, the problem of fixing the size of the reference class against which a hit/miss ratio can be calculated? Machery attempts to bypass the denominator problem using standard sorts of ampliative inferences from properties of observed samples to properties of larger populations, which don't require knowing the overall population size. And he's right that we don't strictly speaking need to be able to calculate the m for the whole population of philosophical case verdicts in order to calculate the ratio n/m provided that we have a good inductive sample to work from; all we'd need in that case is to calculate the n/m in that sample and then deploy standard sorts of ampliative inferences from properties of observed samples to properties of larger populations. Sample enough verdicts in the right kind of way, and find enough problems with them, and you can infer that the verdicts on the whole are equally problematic, no matter how many there are. In this context, we see Machery talking a lot about *typicality*, to provide support that he's considering an appropriate sample; here's one example:

> The first approach is to hold that the philosophical cases examined by experimental philosophers are typical of the kind of case philosophers use, and that their typicality justifies the conclusion that many philosophical cases elicit disagreement. A philosophical case is typical if and only if it possesses many of the properties many philosophical cases possess. For this reason, everything else being equal, typicality is a good cue for induction, and atypicality a good cue that

> induction should be avoided. Everything else being equal, it is reasonable to induce that some bird possesses a property if sparrows possess it, unreasonable if penguins possess it. While it is admittedly not clear what properties make typical cases typical, typicality can be recognized even when such properties are not made explicit, exactly as we can recognize a typical dog or a typical bird without knowing what makes them typical. There is little doubt that experimental philosophers have studied some of the most typical cases, but if you are inclined to doubt this claim, just ask a random sample of philosophers to list philosophical cases, and you'll find, I predict, that the cases examined by experimental philosophers will be the most commonly cited. (p. 109)

Machery's argument here seems to be something like the following. A sufficiently large proportion of experimental studies on philosophical case verdicts demonstrate that those verdicts are unreliable, and since these studies focus on the kinds of cases typically used in the method of cases, we can infer that the method of cases is itself generally unreliable.

This sort of inference is, of course, not deductive, but Machery is not especially concerned:

> supposing that the philosophical cases experimental philosophers have examined are really typical, typicality is at best a fallible cue for induction. Blue jays may have many properties many birds possess, but few other birds transport the nuts and acorns they collect for thousands of meters, and inducing gathering behavior on the basis of their typicality would often lead to mistakes. It may thus be that, while typical philosophical cases—as revealed by experimental philosophers' research—result in unreliable judgments, this is not the case of philosophical cases in general. (p. 109)

Why is he not particularly disquieted by this fact? Because as any philosopher of science will tell you,

> induction is a risky inference. So, while the typicality of philosophical cases cannot guarantee that most philosophical cases elicit unreliable judgments, it could hardly be an objection that inferring on the basis of typicality is fallible. (p. 109)

We, in contrast, find ourselves rather more disquieted. The problem seems to us to be much worse than simply some case where a well-performed inductive inference is suffering perhaps from a mild case of localized fallibility. Instead, we are worried that the basic conditions for the kind of generalization Machery needs here are simply not met. Here's why. Typicality, as defined in the passage above, is a property of *individuals*: *a* is a typical *F* if and only if *a* possesses many of the properties

possessed by a majority of *F*s, especially where those properties are highly distinctive of *F*s. But successful generalization from a sample to a population depends on the properties of the *sample*: namely, that it, as a collection of *F*s, be representative of *F*s at large. And properties of samples do not supervene over the properties of the individuals that comprise them. There are various techniques for ensuring representativeness, such as sample matching, but the classic approach is to gather one's sample randomly; after all, there's a reason Machery tells us in the passage quoted to "ask a random sample of philosophers to list philosophical cases." The typicality of the members of a sample in no way makes it more or less likely that the sample on the whole is representative. We suspect that the bird example may have led Machery astray, since natural kinds provide a much stronger basis for generalization from individuals than other sorts of kinds do. Nonetheless, surely an ornithologist who wanted to publish a paper about, say, the entire population of birds of North America, but generalized from a sample that included only jays, sparrows, and robins—all perfectly typical birds—would surely face some serious negative feedback and not just from a hyper-persnickety referee #2.

It's not just that we have no good reason to expect that the sample of philosophical cases that experimental philosophers have written about will be representative of the kinds of cases used by philosophers. We have good reason to suspect they are *un*representative, biased problematically in favor of error-prone sorts of cases. In our discussion of Knobe in Chapter 2, we appealed to just how vast the space of possible effects is, if we were to randomly mix and match different philosophically irrelevant variables with different target cases. There are at least two reasons to suspect that the sample from that space as reported in the literature is skewed at least somewhat toward "interesting" effects, though we will not speculate as to the total size of that bias: first, publication bias, and second, the simple fact that experimental philosophers are reasonably good at selecting materials that yield interesting effects. We expect there are any number of cases out there that experimental philosophers have examined, but never discussed, for the very simple fact that they didn't find anything interesting to discuss. Experimental work on at least some of these cases is very likely sitting around unpublished in the proverbial file drawer, and while this is no reason to question positive findings about heterogeneity and instability and inconclusiveness that have been reported, it does provide a reason to question just how representative a proportion of philosophical cases work published by experimental philosophers can be expected to be. (Our own (2007) paper with Stacey Swain on order effects grew out of an initial attempt to find an effect of individualism and communitarianism on participant's verdicts on the Truetemp case—a null result that has never hitherto been reported!) What's more, experimental philosophers, like any good empirical researchers, haven't been conducting their investigations at random, but have instead selected their materials generally with an eye toward where we might suspect that interesting effects will be found. This means that, if experimental philosophers have a decent

nose for where to find interesting effects, and if they have focused their inquiries accordingly, then we should expect that the kind of work published by experimental philosophers will constitute a sample more rife with interesting effects than is philosophy at large.

We think that all of this is enough to highlight the biggest concern that we have with the argumentative strategy that Machery takes, namely, that it is really hard to establish that the method of cases is unreliable using anything like the kind of work done by experimental philosophers. Experimental studies tend only to show, at best, that this or that kind of scenario—indeed, this or that specific version of this or that kind of scenario—displays such and such philosophically problematic sensitivity or fails to yield the verdict that is the consensus in the published literature. If all we have are such findings, even rather a lot of them, we nonetheless cannot add them up to get anything like the sort of ratio of total hits to total attempts that constitutes an unreliability thesis.

If we are right, and the kinds of work done by experimental philosophers won't really help us establish anything like Machery's strong unreliability thesis, what can it help us establish in terms of reliability and unreliability, or at least hope to? A couple of things are worth pointing to at this point. First, it seems that the kind of work that has been done by experimental philosophers is already adequate enough to make it a live empirical possibility that the method of cases is, at worst, *moderately unreliable*, that is, a possibility that "for all we know" is true, and worth serious empirical investigation, while in contrast, the sorts of trends we saw Nagel discussing in Section 1 strongly suggest that full-blown unreliability is no longer on the table. Although we aren't inclined to build too much of a case against armchair philosophy based only on the premise that moderate unreliability is a live possibility, we can see how others might try to do so, and we would leave the possible development of that kind of argument to them. More than that, it seems to us that the kinds of studies that have been conducted by experimental philosophers suggest that the method of cases is, at best, moderately reliable and that there are surprising and philosophically problematic pockets of unreliability. We are using "moderately reliable" here as a bit of a term of art, something like being basically reliable but just not too much more reliable than the levels of "moderate unreliability." This will turn out to be an important notion for us moving forward, especially in the next chapter, because an upshot of a method's being moderately reliable is that researchers cannot unquestioningly rely on such a method in a transparent way. We cannot just take its deliverances, and then not bother with keeping track of where they came from. We can rely on such methods but only if we are also keeping ongoing track that that method was the source for them, and under what conditions, and keeping a weather eye out for whether we learn that those conditions place the result in a hitherto undocumented pocket of unreliability. It is important to keep in mind that even such moderately reliable sources of evidence really can be put to good evidentiary use, at least in principle and very often in practice.

Consider, for example, the kinds of survey instruments typically used in social psychology. These are *very* noisy ways of measuring psychologically interesting facts about people, but when aggregated properly, they can still be used to produce good scientific results. In fact, even sources that operate only slightly above chance can sometimes be put to good use under the right kind of circumstances, with the right averaging techniques and so on; yet another reason not to try to use moderate unreliability as a reason to reject a whole source of evidence outright.[14] We see no reason at this time to think that the method of cases is so unreliable that it could not in principle be put to some good methodological purpose—though it may be true, as we will argue in the next two chapters, that it is unacceptably risky to try to pursue such purposes within the dangerous confines of the armchair.

3 Baz and the Unreliability of Both Armchair and Experimental Philosophy

To complete our case against framing these methodological debates in terms of general reliability, we will consider a recent "pox on all houses" argument from Avner Baz. In his 2017 book, *The Crisis of Method in Analytic Philosophy*, Baz argues that both the armchair method of cases *and* experimental philosophy, at least the variety that involves studying the kinds of cases that philosophers use when they use the method of cases, are unreliable methods.[15] Up until now we have, by design, spoken very blandly and neutrally about our "case verdicts," aiming to avoid getting entangled in any but the most ecumenical commitments about the sorts of cognitive processes and mental states might be involved in it. But, Baz cautions, one of these commitments, which he calls "the minimal assumption," is both broadly shared and dangerously false:

> The theorist describes (or otherwise invokes) a "case" and then asks by means of perfectly familiar words a question that has the general form, "Is this (or would such a case be) a case of 'x'?" and gives his answer to that question or collects the answers given by others; or else he simply theorizes on the basis of some tacitly assumed answer to some such question. In proceeding this way, he assumes that the question has a clear (enough) sense and may be answered correctly or incorrectly. He also assumes that, as competent speakers of the language, we—that is, his audience—ought at the very least to understand the question and be in a position to answer it correctly, just on the basis of our familiarity with the words and

[14] E.g., if the conditions of the Condorcet jury theorem are met, and the sample is large enough. See Weinberg 2016b for further discussion.

[15] In addition to those aspects of the book that we will focus on in this section, readers might also be interested in the worries that Baz has about the "arguments not intuitions" move that we discussed in Chapter 2.

> with the case as he describes it. This is what I have called "the minimal assumption." The assumption, as we have seen, is shared by all of the parties to the recent debates concerning the method of cases. (p. 68)

Here is Baz's initial gloss on the *minimal assumption*, where we will italicize a portion of it that we will be giving some critical attention to shortly:

> The minimal assumption is that the theorist's questions, as presented in the theorist's context, are, in principle, in order—in the simple sense that they are clear enough and may be answered correctly or incorrectly—and that, as competent speakers, we ought to understand those questions and be able to answer them correctly, *just on the basis of the descriptions of the cases and our mastery of the words in which the questions are couched* . . . (p. 6)

If both the armchair methods and experimental philosophy make this problematic assumption, then the optimistic spirit that we intend for this book would prove horribly misplaced. The method of cases, in both armchair and experimentally improved forms, would be founded on a strong and seemingly indefensible presupposition. How can we hope to do so much positive and novel philosophy just based on the mastery of our words alone? And for that reason, it is important to respond to Baz here.

The first thing that we want to point out is that we agree with Baz that the minimal assumption, as fully stated, is contentious. This is why it is dialectically useful for him to attribute it to his targets, but it is also why we should wonder whether some even more minimal—and hopefully thereby less controversial—assumption could be identified that could do the same kind of work for licensing practices of appealing to verdicts, either from the armchair or mediated by experiments. We think that there is. In essence, we propose to split Baz's minimal assumption down the middle, in an attempt to excise the noxious portion and keep the philosophically near-platitudinous remainder. Our proposal is this. We think that everything before the italicized portion of the passage that we just quoted must indeed be a shared basic ground for both armchair and experimental philosophers. As we might alternatively phrase it: the cases themselves must have verdicts, and those verdicts can be right or wrong, and practitioners must be able to understand the cases well enough to render a candidate verdict, and they must be able to report in a publicly available manner just what verdict it is that they have arrived at. But the bit we have marked in italics contains all the hard to swallow philosophical gristle here—and as such ought to be left on the methodological butcher's block. We agree that philosophical and experimental participants alike must be competent enough to be able to comprehend, judge, and report their verdict on the relevant philosophical cases. But why think that this competence must be based *solely* on the descriptions, in isolation, combined with the bare linguistic mastery of the

terminology therein? We can think of no good reason for adding these demands, and so want to distinguish between the unitalicized section, which we might call the *even more minimal assumption*, and the italicized section, which we might call the *distasteful addendum*. We are willing to grant to Baz that the (not-so-)minimal assumption, so long as the distasteful addendum is left in, should be a source of philosophical indigestion. What we want to argue is that there is a range of other ways of characterizing and valorizing our shared categorizing competences, in ways that give both armchair and experimental verdict-mongers what they need, without the problematic commitments Baz looks to hang upon them.

First, it is not remotely a consensus view that when people think about philosophical cases they should do so without drawing on various sorts of information to arrive at verdicts about those cases beyond the kind of information that is captured in their comprehension of the language used to describe the case. It does seem to be the case that some philosophers who have been interested in defending the method of cases, and especially the armchair-bound version, have a vested interest in case verdicts being *a priori* or analytic, and, for this reason, may need to take on board something like the distasteful addendum that Baz includes in his minimal assumption. But experimental philosophers, as we discussed in Chapter 1, are not at all likely to take anything like this on board, since they are often interested in learning specifically about what kinds of things *other than the kinds of information captured in linguistic competence* influence people's verdictive capacities. Perhaps this is because they are trying to make trouble for the armchair-bound version of the method of cases, or simply because they want to make sure no such effects are cluttering up their main substantive results. What's more, a significant number of armchair philosophers have been happy to accept that people's case verdicts reflect a host of information beyond the kind of information that is captured in their comprehension of the language used to describe the case; as we also noted in Chapter 1, many construals of "armchair" are happy to include various sorts of empirical information within our upholstery. And considering our articulation of the praxis of the method of cases in that chapter, we see nothing like the distasteful addendum built into that method itself, certainly not explicitly but also not, so far as we can see, implicitly.

We suspect that part of what has gone wrong here is that Baz, taking Williamson as his stalking horse as a defender of the method of cases, attributes to him a restriction to "mastery of the word," when he talks about what's involved in forming the kinds of counterfactual judgments that he thinks are characteristic of how people think about philosophical cases. Baz quotes the following sentence from Williamson's *The Philosophy of Philosophy*: "We assent to the Gettier proposition on the basis of an offline application of our ability to classify people around us as knowing various truths or as ignorant of them" (Williamson 2007, 188, quoted in Baz 2017, 74). Baz interprets this "offline application of our ability to classify" as involving a restriction to our bare linguistic competence. Yet Williamson seems to

us to mean, in contrast, a wide range of empirically acquired information—hence his stark rejection of the *a priori/a posteriori* distinction.

We suspect that the other part of what has gone wrong here has to do with Baz's own philosophical commitments. In particular, Baz seems to think that "theorists" are committed to the minimal assumption precisely because he endorses a radical contextualism about language that, it seems to him, is inconsistent with the sort of isolated, lepidopterist-pinned way of presenting philosophical cases that does indeed seem to be part of both traditional and experimentalist practices. In a nutshell, Baz worries that our competence to apply or withhold application of our words, or, importantly, to recognize that a proposed scenario evades that competence, is utterly contextual, in a sense that goes beyond increasing the arity of the knowledge predicate with a "context" slot. The ambitious nature of Baz's form of contextualism is usefully glossed in a review by Nat Hansen (2018), who also has some nice discussion of ways in which something very like contextualism is already a matter of explicit concern and interest in the social sciences:

> More controversial types of contextualism claim that the truth conditions of what is said by the use of sentences containing certain expressions which aren't obviously indexicals, like 'know' or 'might' or 'good', are not fixed by the conventional meaning of those sentences, but can vary in different contexts in more or less unobvious ways. 'Radical' contextualists maintain that the conventional meaning of a sentence (with the possible exception of mathematical sentences) is never sufficient to fix the truth conditions of what is said by a use of that sentence. (pp. 964–965)

In other words, Baz thinks that we cannot even talk meaningfully in any sort of general terms about reliably rendering verdicts about philosophical cases, because these cases by themselves do not generally induce application conditions for key pieces of philosophical terminology. We cannot be right in a large majority of the cases, because there isn't (yet) anything to be right about, when such cases are disembodied chunks of language stranded from any sort of real, lived human communicative situation.

We don't want to argue here against such radical contextualism, and so will remain neutral about such theories about language. Instead, we want to articulate a way in which appeals to case verdicts would most likely be consistent with such radical contextualism in the first place: after all, *the theorist's context is a context in its own right*, and we can consider whether that context can do the requisite work to help fix correctness conditions for verdicts. In order to do this, we need two pieces of machinery, which are not beyond philosophical dispute, but which we think, nonetheless, are fairly uncontroversial. First, we need a version of the representational theory of mind, and in particular a version that gives us a fair amount of stable, context-independent mental representation. We are confident that some

such theory is, in fact, presupposed by radical contextualism, since otherwise it would be a brute mystery how human cognizers can compute the relevant particulars about whatever communicative context they happen to find themselves in, in order to determine what might be said and done by some candidate utterance. There's probably some sort of transcendental argument to be made about how any serious amount of contextualist linguistic machinery presupposes substantial context-independent machinery: the more complicated a story one wants to tell about linguistic performance, the more sophisticated the mental equipment one must invoke in order to implement those complexities. We will remain officially agnostic here as to what sort of structures or processes may be needed to operate over those representations; in particular, they may not need to be linguistically structured like a Fodorian "language of thought." And for the sake of argument here, we will stipulate that there does not need to be any simple, transparent, context-independent mapping from mental representation to linguistic expression. It is thus, importantly, consistent with Bazian radical contextualism about linguistic comprehension and production.

The second piece of machinery that we need concerns how our mental processes operate when considering a hypothetical scenario. Presumably we do this by tokening, in the imagination, representations of the contents of the scenario. Importantly, we almost always entertain not just the explicitly stipulated contents in the text, but go on to substantially expand on that content, filling in with all sorts of information from our stock of background beliefs. Of particular relevance here, we have good reason to think that folks considering standard thought experiment vignettes will be automatically and effortlessly calculating whether agents in the story do or don't know various salient propositions, what events in the story are or aren't causes of other events, whether various actions are morally permissible or impermissible, and so on, regardless of—and indeed, prior to—any explicit probe question about such matters. Such calculations are made all the time without any prompting, and may also be made immediately and with little to no conscious effort in response to a prompt, either in the story itself (if, e.g., another agent makes an explicit knowledge attribution to someone), or externally (if, e.g., issues of knowledge are raised by the philosopher or psychology "theorist").[16]

These two pieces of machinery are enough to get to merely considered verdicts, affirmed in the mind but not yet expressed, but our preferred even more minimal assumption needs to get us all the way to some sort of reports that are amenable to the "answer them correctly" sub-clause. The last key move, then, in offering a Bazian-contextualism-friendly theory of thought experiments is to note

[16] See, e.g., Van Leeuwen (2013) on what he terms the "Belief Governance Thesis"; his "Genre Truth Governance Thesis" is also perhaps salient to the next paragraph. For some discussion about the role of such filling in for our engagement in thought-experiments, see Weinberg (2008), Camp (2009), Ichikawa and Jarvis (2009), Sosa (2009), Alexander (2012), and Powell et al. (2013).

that, among our many culturally learned and contextually bound linguistic competences, is the capacity of just articulating whatever imagined contents we are entertaining at a time. The previous two assumptions ensure that such imagined contents are themselves something that is meaningful to talk about, and that this will include hitherto unstated elements of hypothetical vignettes. Baz, with his distasteful addendum, takes it that we must be assuming that the mastery of the words *all by themselves* is enough to enable this capacity. We want to suggest, instead, that it may well be that this is something that requires significant acculturation and training, and may be itself a context-bound capacity *in just the way Baz wants to insist*, and yet it is a capacity that is extremely widely cultivated in the world today. We offer no speculation as to whether it is universal, especially among preliterate societies. But in our culture, we are trained from an early age to perform well on reading comprehension tasks, math word problems, and the quizzes in *Cosmo* magazine, or those found online to determine "which character are you?" Entertaining these sorts of free-floating on-the-page questions and being able to answer them sensibly is something that we can expect every philosopher to have been trained to do long before their first philosophy class.

For a further illustration, consider our practices with following along with fictions, appropriately filling in unspecified contents, and then being able to report on what we are imagining to be true in its world, if someone just asks us, "hey, what's going on in that story you're reading?," or if we are talking with a friend who is in the middle of the same serialized fiction and we are debating with relish what may or may not happen next, or what mysteries are or aren't yet resolved. It's clear that there's nothing exotic here, even with philosophically important categories. Consider how, if you are discussing a detective story, you will naturally and without special philosophical prompting be keeping mental track of, and reporting on, which characters do or don't know what, even in the absence of any flat-out assertions of knowledge to that effect in the text itself (which in that genre are more likely to be ironically inaccurate, anyhow). What we are suggesting, at least for the purposes of demonstrating the consistency of our even more minimal assumption with Baz's radical contextualism, is that these practices and the capacities that subserve them may well be the product of a highly trained up and contextually bound mastery. (As parents of young children, we certainly can affirm anecdotally that at least sometimes, getting your children to just simply say what they are imagining in their heads is as nontrivial a task as Baz's rejection of the minimal assumption would suggest!) The philosopher's armchair and the psychologist's survey are not utterly contextless, nor must the theorists presuppose them to be.

Indeed, this sort of competence in hypothetical verdict formation and self-report is one that a great many working psychologists presuppose in their participant samples when they do their own studies of how people make judgments about various situations. On Baz's account, when people are asked to offer an allegedly decontextualized verdict about a hypothetical vignette, instead of answering with

a positive or negative verdict, the "proper, correct, response to it would be to say, 'It all depends on what you mean, and that is not something that your words alone could determine'" (2023, 83). But, overwhelmingly, participants in such studies do not do this. Of course, if the vignette is weird enough, they might balk and answer at random, or as close to the midpoint as the particular study design allows. And there is always a practical methodological issue of how the participants might have *more* in mind than your design lets them offer, and so their response on your instrument gets thereby distorted.[17] But the generally smooth running of such experiments shows that, even if we stipulate that Baz is right about the proper response in a truly decontextualized setting, it appears that the theorist's questions are not being asked without context.

Now, the theorists may turn out to be interestingly and importantly *wrong* about elements of such contexts, and about how either trained philosophers or the philosophical laity will go about trying to respond to vignettes. And there remain methodologically thorny issues about determining when, or to what extent, different readers of a vignette are filling in details in a sufficiently similar fashion.[18] But that concern doesn't lead to a radical rejection of case verdicts as evidence, but rather puts us back in the realm of empirical questions about errors and their mitigation.

4 The Method of Cases and the Limits of Epistemic Normativity

We have argued that methodological questions about the methods of cases cannot be answered in terms of reliability and unreliability, and the reason for this has everything to do with the concepts of *reliability and unreliability* themselves. Not only does it turn out to be impossible to show that the method of cases is unreliable using X-phi evidence, but it turns out that when it comes to methodological questions, how reliable our belief-forming practices have to be in order to count as reliable enough turns out to be relative to specific domains and purposes, and there are reasons to think that philosophical practices and purposes are sufficiently dissimilar to the kinds of ordinary cognitive processes and practices that analytic philosophers often appeal to when defending the method of cases. Moreover, it may be that no plausible degree of reliability for case verdicts could be enough given the error-fragility of our philosophical modes of inference. This means that it is not enough for analytic philosophers to argue that the method of cases is a natural extension of more ordinary cognitive processes and then argue that the reliability of these ordinary practices extends to the method of cases. By the same

[17] This is a major issue in much of experimental psychology and X-phi, but it is important to see it as a practical problem to be addressed and overcome, and not a fundamental objection to the cogency of the research. See, e.g., Cova et al. (2016).

[18] Which is also, frankly, itself a methodological issue for a Baz-style ordinary language philosopher as well. Our thanks to Jennifer Nado for raising this issue of imperfectly shared fillings-in.

token, it is also not enough for experimental philosophers to find evidence that such and such case verdicts are problematic in this or that way, and then argue from such evidence to a general unreliable thesis about the method of cases. We will argue in the next chapter that this problem is not peculiar to reliability and unreliability, but indeed that the traditional concepts of epistemic normativity on the whole will prove to be of no real help here. Advancing the debate between the experimental challenger and the armchair defender will require a new normative framework. In the next chapter, we will provide it.

4

A Better Normative Framework: Methodological Rationality

In Chapter 1, we set out the basic contours of an experimental challenge to armchair philosophy. We want sources of evidence, philosophical or otherwise, to be sensitive only to the right kinds of things, namely, those things that are relevant to the truth or falsity of the kinds of claims for which those sources of evidence are supposed to provide evidence. But there is a large and growing body of evidence in experimental philosophy that suggests that our case verdicts are sensitive to more than just these things. They can be sensitive to aspects of who we are and the manner in which we are asked to think about philosophical cases. This kind of unexpected and worrisome evidential sensitivity is made worse by the fact that we know so little right now about philosophical cognition, that is, about how our minds work when we think about philosophical issues. This means that we have no real way of knowing at this time the full scope of the problem. Our current state of knowledge about such effects does not come close to telling us all of the things to which our case verdicts will be sensitive or which case verdicts will be sensitive to those things. As matters stand so far, we take it that the experimental challenger should have succeeded in worrying the armchair defender, yet the latter can legitimately feel that it has not yet been determinately shown that this worry rises to the level of dislodging them from their seat. In the previous chapter, we argued that the standard epistemic normative concepts of *reliability* and *unreliability* do not present a fruitful way to think about this metaphilosophical challenge. This raises the obvious question, what *is* the right way to think about this challenge? In this chapter, we will propose a novel answer to this question, and offer a better conceptual and normative framework for thinking about the metaphilosophical implications of experimental philosophy for armchair philosophy. Our plan of attack is to give something like the traditional epistemological conceptual framework its best possible shot, and when that is seen to fail, we will construct alternative normative machinery, using tools from the domain of practical rationality.

Standing Up for Philosophy. Jonathan M. Weinberg and Joshua Alexander, Oxford University Press.
 DOI: 10.1093/9780191947704.003.0004

1 What's *Really* Wrong with the Standard Normative Framework

Let's start again at the beginning. The most common way these days of thinking about the relationship between experimental philosophy and the method of cases fits what we will call an *impeachment schema*:

> Epistemological Premise (EP): Any putative source of evidence that is β is severely deficient, which means, at a minimum, it is incapable of providing evidence.
>
> Descriptive Premise (DP): The method of cases is β.
>
> C: The method of cases is incapable of providing evidence.

Different philosophers have suggested different properties for β resulting in different metaphilosophical arguments against the method of cases. In Chapter 3 we looked closely at the most popular and influential version of this argument schema, where β = unreliable. It is worth noting, in the context of the concerns that we raised there about that particular version of this impeachment schema, how hard it actually is to find a good candidate for β, especially since EP and DP create a Scylla-Charybdis dynamic that has proved hard to navigate.[1] Set the bar for β too low and DP becomes really likely to be true but EP becomes really likely to create unwanted headaches since lots of other sources of evidence will likely run afoul of it as well, such as ordinary perception or even scientific methods.[2] As we saw toward the end of Chapter 1, Sosa warned about such a skeptical result as a clear consequence of setting β for the too-low bar of being *merely fallible*, since sense perception suffers from all sorts of fallibilities similar to those that X-phi has documented about case verdicts. Conversely, if we set the bar for β too high in order to avoid such untoward consequences, we might keep it from applying to sources like ordinary perception—but only at the risk of DP being false or, at least, unlikely, and highly contentious. Remember that one of the takeaway points from Chapter 3, where we considered β = moderately unreliable, was that there seemed no good way available to try to argue that DP was actually true (and not just an epistemic possibility) given that parameter setting.

And this is just to merely dip our oars into the surface of these dangerous waters. We could probably read the entirety of *Odyssey* book XII before we could finish sorting all the various failed candidates for β into whether they are EP-falsifying or DP-falsifying, but here we will offer just one last example from our own perilous voyages in these debates. In response to Sosa's admonition about being careful, it always seemed to us that there is, in fact, an important disanalogy with ordinary

[1] See Kumar and May (2019) for a useful discussion of this kind of problematic dynamic in the case of moral judgments.

[2] The shoals of self-refutation also lurk in these skepticism-infested seas. One interesting side-debate has focused on whether experimental philosophy risks self-impeachment. We will take the fifth, but for discussion see Horvath (2010) and Machery (2017).

perception. And so, in the past, we used the term *hope* as a technical term for the resources that members of a research community have available to help them carefully use the methods that are central to their research (Weinberg 2007, Alexander 2012). The fallibility of a hopeful research method could be mitigated on the assumption that researchers take the measures that, *ex hypothesis*, they know how to take. If a hopeful method, in this technical sense, is one where researchers know how to be careful with it, it follows that the technical term *hopelessness* will apply to methods of *unmitigated fallibility*, "fallibility uncompensated by a decent capacity for detecting and correcting the errors that it entails" (Weinberg 2007, 323). It's important to see how hope, while still being a veritistic virtue, is conceptually quite distinct from reliability. Where reliability is all about maximizing the chances that our methods get things right in the first place, hope is all about how well we can fix things once we realize that our methods sometimes yield mistakes. Accordingly, our methods are (in this technical sense) hopeless when we lack the resources needed to fix these mistakes. This means that it is possible, for example, that our case verdicts might not be particularly worse off than ordinary perception in terms of reliability, but could still be sharply different in terms of hope. Put into our terms here, we argued in those earlier works that hopelessness would be a good candidate β for running an impeachment argument. If philosophers totally lack any resources to detect or correct for errors in the case verdict evidence, then we hoped that perhaps that could be a solid reason to take the case verdict evidence to be severely deficient, without the kinds of skeptical worries that worried Sosa.

Alas, these hopes were also dashed against the rocks.[3] The problem is that hopelessness turns out to be DP-violating, since there are surely a number of errors that we do know how to look out for in existing philosophical practice, even from the armchair, and that we are able to correct when found (see, for discussion, Grundmann 2010, Horvath 2010, Ichikawa 2012, and Osborne 2014). Consider scope ambiguities, or pragmatics/semantics confusions, or deeply hidden forms of begging the question to start a list that could be extended extensively. Analytic philosophers are trained to know that these problems can arise, and even better, to know how to look for them when we think they have. And we appropriately devote resources (such as research time and space in academic journals) to doing so. In short, there are a number of errors that philosophers do know how to be careful about. Now, of course, these aren't the sorts of problems with our case verdicts that experimental philosophers have been revealing, but the fact that philosophers have a well-developed facility for handling them manifests that there is no *global* lack of hope in our practices. A blanket charge of hopelessness just won't stick.[4]

[3] Fans of the Odyssey will notice that this isn't quite the right way to cash out the metaphor. Both Scylla and Charybdis are monsters, after all. We've changed the metaphor here because, while we strongly disagree with our interlocutors, we don't think that they are metaphilosophical monsters! (Not most of them, anyway.)

[4] Brown (2013) argues that hopelessness is also EP-violating, at least in certain methodological circumstances. For example, if a method is at least reliable, but it's all we have to go on, we ought to go

Now, while *completely lacking any ways of detecting errors* makes for a poor impeachment β, what about a more restricted and localized kind of hopelessness? That is, what if our methods have some significant exposure to specific risks of error and lack the resources to monitor or correct for those specific risks even when researchers using those methods are not utterly devoid of ways of detecting and correcting other sorts of errors associated with those methods? This is a particularly promising alternative impeachment β, since the kind of error-fragility that we argued in Chapter 3 attaches to philosophical inferences means that even a few pockets of uncontrolled, unexpected sources of error in our case verdict evidence could render highly suspect any inferences based on those verdicts. So, let's try to build a workable impeachment β out of this rough idea. For ease of discussion, we'll make the following coinage: a research community *C* using method *M* is *vulnerable* to type of error *E* iff (i) *E* is *empirically live* for *M* under the conditions in which *C* is operating; and (ii) *C* lacks adequate resources to *mitigate* the error risk posed by *E*. By "empirically live," we mean a more substantive condition than mere nomological possibility, yet at the same time, something much weaker than its being positively likely that *E*-type errors will occur. If we're looking to construct a β for an impeachment argument, then that first reading would lead to an EP-violation, since, for example, it is nomologically possible that mathematicians very frequently make mistakes in their proof-checking practices, yet we do not mean to be impeaching the method of proof in mathematics. Possibilities, even nomological ones, are too cheap. And the second reading would lead to a DP-violation, as indeed something like a targeted version of a moderate unreliability thesis. Instead, by "empirically live" we will mean something more like: a chance of error sufficiently greater than zero that it must be taken seriously as a practical matter for inquiry. We can look to guidance here from good experimental practice in the sciences, where we see what sorts of error possibilities scientists regularly take measures to compensate for. When someone proposes a possible confound for an experimental result, it can't be merely possible-but-preposterously-unlikely, but at the same time, lots of even fairly remote, and on the whole unlikely, possibilities get taken very seriously by scientists all the time, and appropriately so.

It is clear that the first clause, that is, that *E* is *empirically live* for *M* under the conditions in which *C* is operating, would make a lousy candidate for an impeachment β because even our best scientific methods often leave open these kinds of error possibilities. Such error possibilities can often not be eliminated, but only managed. It is the risks that we cannot manage, or maybe do not even know that we need to manage, that pose a more truly dire threat to inquiry; this is why the second clause must be added, to the effect that *C* lacks adequate resources to *mitigate* the error risk posed by *E*. We need to be clear that we are refining our terminology a

ahead and do our best with it, no matter how hopeful it is or isn't. We will consider an adapted version of this argument in the next chapter.

bit here, from thinking about hope in terms of our specific capacity to detect and correct for possible errors to the broader set of actions that all count toward mitigating the risk posed by a source of error. We are doing this primarily because very often the risk of error is managed not by catching specific errors after the fact, but by adopting other measures to shield our inferences and theories from harm. For example, social psychologists may not be able to eliminate order effects from some particular survey instruments, but they are able to compensate for those effects by presenting their survey instruments to different subjects in different orders, and then averaging across those different conditions; whatever effects are still observed are not subject to any unwanted influence of order. Or, when natural scientists are trying to fit a curve to a set of observations, they may well know that the instrument being used was highly noisy and that any individual datum is likely to be somewhat inaccurate. Hope, as originally phrased, would involve something like being able to go back and take a new set of measurements, perhaps with different instruments or at least making independent trials with the original one. But even if an observational do-over is not practical or even possible, measures are still available to minimize an erroneous choice of model: if they know their data are noisy, then they know that *ceteris paribus* they should prefer to trade off cleaving too closely to the data in favor of a simpler curve. Under such circumstances, if they stick too close to the particulars of their data, then they are likely over-fitting and mistakenly carrying the noise along in their inferences when they should be filtering it out.[5] In short, the means of error management are myriad and sundry, and which ones to use where and when surely depends enormously on utterly local factors. But when we can't do it at all, we are likely in deep methodological trouble.

With all of this in mind we can evaluate whether *unmitigated vulnerability to specific risks of error* is a good candidate for an impeachment β. One consideration in favor of using something like this as our impeachment β is that it seems to split the difference between the too-low bar of mere nomologically possible fallibility (that is, just clause (i) by itself) and the too-high bar of general moderate unreliability (which we saw in Chapter 3 is unworkable). Moreover, this candidate for an impeachment β yields an extremely plausible and likely version of the descriptive premise. As we discussed in previous chapters, we have (and continue to gain) very good reasons at this point to think that the method of cases is exposed to a significant number of empirically live sources of error. As we have emphasized, the kinds of problematic sensitivities that experimental philosophers have discovered fall, by and large, outside the observational resources of the armchair. They are too subtle to be noticed by unaided and unsystematic observation, too unconscious to be revealed by introspection, and often fall outside of armchair experience altogether (as many group differences will).

[5] We discuss such inferences further in Chapter 7.

While there are, therefore, reasons to think that *unmitigated vulnerability to specific risks of error* is a good candidate for an impeachment β, we will nonetheless *not* be advocating this property for the impeachment schema because, as a candidate for β, it seems to us to violate the epistemological premise. When substituted into the schema, the resulting claim just does not strike us as true:

> UV: Any putative source of evidence for which there exists an unmitigated vulnerability to error is severely deficient, which means, at a minimum, it is incapable of providing evidence.

The core problem is that while a method with unmitigated vulnerabilities clearly has *some* sort of serious troubles, those troubles do not seem to us to add up to an utter incapability to provide *some* amount of evidence. If the source is moderately reliable, then it is carrying significant information about its targets—it has some very real signal in all of the noise—and UV thus fails to solve the Scylla-Charybdis problem by committing a version of the baby-bathwater one.

There is an argument to be made here from the history of science. For example, we do not think that all empirical work done prior to modern double-blinding protocols is in fact completely devoid of evidential value. The same thing can be said about mathematical work done before the full development of modern proof-checking practices. To take just one more, slightly more in depth example, consider Chang's (2004) extraordinarily enlightening discussion of the development of thermometers. Here's a particularly helpful passage for the point that we want to make here, where Chang talks about early work by Daniel Fahrenheit:

> Fahrenheit made some important early experimental contributions to the study of specific heats, by mixing measured-out amounts of fluids at different initial temperatures and observing the temperature of the resulting mixture. In these experiments he was clearly aware of an important source of error: the initial temperature of the mixing vessel (and the thermometer itself) would have an effect on the outcome. The only way to eliminate this source of error was to make sure that the mixing vessel started out at the temperature of the resulting mixture, but that temperature was just what the experiment was trying to find out. The solution adopted by Fahrenheit was both pragmatic and profound at once. In a letter of 12 December 1718 to Boerhaave, he wrote:
>
> > (1) I used wide vessels which were made of the thinnest glass I could get. (2) I saw to it that these vessels were heated to approximately the same temperature as that which the liquids assumed when they were poured into them. (3) I had learned this approximate temperature from some tests performed in advance, and found that, if the vessel were not so approximately heated, it communicated some of its own temperature (warmer or colder) to the mixture. (van der Star 1983, 80–81)

> I have not been able to find a record of the exact procedure of approximation that Fahrenheit used. However, the following reconstruction would be a possibility, and would be quite usable independently of whether Fahrenheit used it himself. Start with the vessel at the halfway temperature between the initial temperatures of the hot and the cold liquids. Measure the temperature of the mixture in that experiment, and then set the vessel at that temperature for the next experiment, whose outcome will be slightly different from the first. This procedure could be repeated as many times as desired, to reduce the error arising from the initial vessel temperature as much as we want. (pp. 225–226)

To take Fahrenheit to lack evidence altogether for various specific heat claims while he was working out this procedure would render unintelligible this, in fact, highly clever piece of scientific research. We might somewhat downgrade the evidential weight of instruments and methods that we know are subject to errors we cannot, right now, completely eliminate, but it doesn't seem that past inquirers have generally considered them evidentially bankrupt, nor does it seem we should do so now. As the history of methodology progresses, we don't always, or even often, throw all extant results and methods overboard and start anew. Rather, we often re-examine previous work with a newly critical eye, and perhaps try to see whether the results can be re-established using methods that are not so vulnerable, or at least where those continued vulnerabilities are mitigated to a greater degree. A transition in our methods from unmitigated to mitigated sources of error should spur a period of much separating of wheat from chaff, using those newly available resources for error mitigation. Such a threshing project should nonetheless not be confused for burning fields to the ground.

We think that the lesson to be learned from all of this failed β-testing is that it has been a mistake all along to look for some feature of the method of cases that is so epistemologically problematic, or otherwise epistemologically objectionable, that it falls unrecoverably beyond the methodological pale. As we said in Chapter 1, we think that experimental philosophy's dirty little secret is that the method of cases is probably not in such a bad way that it cannot *sometimes* be put to good use *somewhere* and *somehow*. This is why we think that it has been so hard to come up with any defensible and compact impeachment β that applies to the method of cases that would make both EP and DP even plausibly true. What's even more important, there are lots of things that researchers can do to make good use of even moderately reliable methods by means of mitigating the vulnerabilities to error. As we observed above, mitigation comes in many forms, and we will offer a diverse range of concrete proposals in Chapter 7 for how philosophy can do better in this regard. But for now, our point is just that it is just really unlikely that we will ever be in a position to make the kinds of sweeping and damning epistemological generalizations about the method of cases that have become such a standard part of the current debate about that method.

In short, the impeachment schema is a poor frame for the experimentalist challenger to make methodological trouble for the armchair. That schema tries to set up a lethal epistemic chop, but the experimental results only yield at best a rather more delicate piece of cutlery than the executioner's axe. What is needed, then, is an argumentative framework that, while it will not enable a full-scale decapitation, can set a methodological operating table for life-saving, or at least inquiry-saving, surgery. The problem with UV is not so much that it has misdiagnosed the illness, in terms of unmitigated vulnerability to error, but rather that it seems only able to prescribe an evidential euthanasia as a course of treatment.

Part of what is tricky in moving away from the frame of the impeachment schema, however, is that it is hard to express and defend any methodological *urgency* in traditional epistemic terms without making the stakes as sharp as something like severe epistemic deficiency. We think the history of inquiry, up to and including the present day, suggests that well-functioning research communities take concerns about vulnerability seriously. When their practitioners become aware of vulnerabilities to error, they feel, correctly, that as a community they ought to do something about it; and most often, they do ultimately identify positive measures to bring their methods into a state of mitigated risk. But how can we convey that normative seriousness, that "ought," in traditional epistemic terms? We have already seen the inadequacy of reliability talk in this regard. Moreover, if we try to operate in conceptual terms like the possession of knowledge or full justification, then the most natural tools to work with to represent such seriousness would all involve setting necessary conditions, and then we would have to argue the method of cases does not currently meet those conditions—which is to say, something like the impeachment schema. But if we, instead, argue in more gradient terms, like degrees of justification or evidence, then we will have trouble establishing a conclusion that anything much needs to change. "We'll grant," offer philosophers who want to defend our armchair practices, grandly, "that the experimental results indicate that we have a bit less evidence than we thought we did. Well, then, all we would need to do is just pause for a moment to turn the knob on our philosophical credences down a smidge or two, and then we can get right back to work the same way we have been. It's not like anyone today thinks that we have geometry-proof-level evidence for our philosophical theories anyhow." This is clearly a recipe for stasis, not reform.[6] And as such, this dialectical move

[6] And it is perhaps worth saying that analytic philosophers would probably be *right* to reply to such an argument in that sort of way. There are no principles that say *always do whatever you can to increase your degree of justification for believing that p* or *always do whatever you can to increase your degree of justification for believing that p so long as you do not know that p* or *always do whatever you can to increase your degree of justification for believing that p so long as your justification is below level* ε. As we will argue in the next section, you simply cannot get from "I have a lower degree of justification than I thought about *p*" to any *particular* course of action, without either using some sort of threshold, or bringing in a broader framework of practical rationality, in terms of the costs and benefits of various courses of action that might be available (including the action of making no changes at all).

would badly overgeneralize, rendering inscrutable the normative motivations for the many substantial, well-merited, and productive methodological changes that make up the history of inquiry. For example, we would not have wanted, say, mid-twentieth-century psychologists to make such a purely defensive and retrenching move in the face of evidence of experimenter effects. It was clearly methodologically appropriate, indeed mandatory, that they made the changes to their practices that they did, to mitigate their exposure to that source of error. Yet it seems we cannot capture the strength of that need for change in the traditional epistemological vocabulary of knowledge, justification, degrees of evidence or confirmation, and so on. Demanding a lower alpha and more participants or something like that simply would not go any distance toward solving this problem.

Another sign that the vocabulary of epistemic normativity may be inapt for framing these issues is that it seems to contribute to one ongoing pathology of the dialectic of experimental philosophy versus analytic philosophy: namely, the tendency of analytic philosophers to rush to treat the kinds of arguments made by experimental philosophers as forms of skepticism. We saw something like this error when we talked about Nagel's argument for the general reliability thesis in the preceding chapter. We also see it in the way that Williamson seems to think that his rejection of "judgement scepticism" in the *Philosophy of Philosophy* constitutes an adequate defense against the worries that experimental philosophers have raised about the methods used by analytic philosophers. While in theory our epistemological theorizing is infinitely nuanced, as a more practical matter in the actual methodological debates, the evidential status knob on the epistemology machine just seems to lack any substantial settings between "skepticism—reject utterly" and "basically fine—infer at will." (Maybe with an epiphenomenal signal light saying, "Be careful!")

The way forward, we think, is a wholesale abandonment of the urge to think about the method of cases in anything like the conceptual framework of *epistemic* normativity, and instead to bring to bear some version of *practical* rationality.[7] The sort of urgency we have in mind is, after all, a practical one, about *what ought to be done*. Of course, the kinds of actions in question are ones with a distinctly epistemic flavor. Are we conducting our philosophical inquiries in the way that we ought? Should we instead be doing something different, or at least substantially differently, from what we are doing now? But they are no less practical for all of that. What we want to do, then, is to use an epistemically inflected form of practical rationality to treat the concerns that have been raised by experimental philosophers about the method of cases, and to use the term *methodological rationality* to label it. We don't mean for this to be a *sui generis* form of rationality, but instead a way of thinking about philosophical inquiry in the instrumental terms of

[7] We are inspired here by Cohen (2012), although perhaps not in a way in which he would approve.

what results philosophers are trying to achieve, and what resources and plans of actions we should adopt in order to maximize how well we can achieve them. The goals in question are the goals of inquiry, and so we have in mind here a special case of practical rationality that can be applied to different modes of inquiry, including the method of cases, in both armchair and empirically inflected forms. In the specific context of philosophical inquiry, methodological rationality is not directly concerned with questions about the epistemic rationality of our philosophical beliefs, for example, whether our credences are the ones that are mandated by the evidence we possess. Instead, methodological rationality applies the general means-ends rationality of practical reasoning to the goal-directed activity of philosophizing, as constrained by the limited resources we may have in pursuing our philosophical goals.

It is worth noting that while there is significant philosophical debate about whether epistemic norms, themselves, are ultimately grounded in terms of practical rationality, we do not wish to take a stance on those questions here. But even those who argue for the independence of epistemic norms acknowledge that there are forms of instrumental rationality that apply in epistemically fraught circumstances. For example, here is Thomas Kelly, in his (2003), arguing against instrumentalism about the epistemic:

> On anyone's view, the fact that I possess certain cognitive goals can make it instrumentally rational for me to do things which it would not be instrumentally rational for me to do, if I did not possess those goals. Suppose that, wanting to know the identity of the person who committed the crime, I engage in the activity of looking for evidence which bears on the question. Here, the fact that I have the goal of learning a certain truth gives me an instrumental reason to act in a certain way: all else being equal, it is rational for me to engage in the activity of looking for evidence. Uncontroversially, the rationality in play is an instrumental rationality in the service of a cognitive goal.

This uncontroversially in-play application of instrumental rationality provides precisely the kinds of terms into which we want to reconfigure the current debate about the method of cases.

Jennifer Nado (2017) has raised similar doubts about the appropriateness of traditional epistemic terms for arguing about the metaphilosophical implications of experimental philosophy. She cautions that

> experimentalists and defenders of intuition are, in fact, frequently talking past one another; while defenders busy themselves with arguments about intuition's ability to justify belief or to generate knowledge, such defenses fail to address the real challenge that experimental findings potentially present to our current methodological practices. Ultimately, all this points to a more general moral: The

> monolithic focus on 'knowledge' as the primary epistemic state of interest in philosophy obscures many of the subtleties involved in evaluating our multifaceted practices of inquiry, both in philosophy and in other fields. There exist many different epistemic standings of interest; it is, I think, a losing game to attempt to capture all these with any reasonably unified account of knowledge. Worse, it is a distraction from the more central question of how to best investigate reality. (pp. 146–147)

And it's not just the everyday notion of *knowledge* that is the problem here, but the whole network of ordinary epistemological concepts. Her advice is that experimental philosophers should "separate methodology from epistemology; instead of claiming that intuition fails to yield justification or evidence, experimentalist critics might instead claim that intuition (at least as currently used) fails to meet the rigorous methodological standards to which philosophers ought to ascribe" (p. 164). We agree with this sound advice, but disagree with Nado about how best to follow it. She writes in terms of professional inquiry, be it philosophical or scientific, typically having *higher standards of reliability* than ordinary standards for knowledge and the like. Thus, on her account, we still use something very similar to traditional notions, but of a more demanding sort—not just, say, reliably produced true beliefs but *very* reliably produced true beliefs. But, along similar lines to our disagreement with Nado in the previous chapter, we don't think that the right way to come at this issue of methodological rigor is to focus exclusively on increasing reliability. As long as we are still talking about greater or lesser degrees of moderate reliability, then there are going to be live threats of error, and potential vulnerabilities, and those threats still need to be included in our working theories about the methods and instruments. That is what we plan to do. By shifting to methodological rationality, we set aside questions like: can the armchair method of cases produce knowledge or justified belief in its practitioners? Instead, we will pose the question: in terms of the goals of philosophical inquiry, what does a cost-benefit analysis of staying in versus leaving the armchair look like? Would the potential benefits in terms of error-management offset the costs of rising to our feet?[8]

[8] Two last notes of caution. First, issues of methodological rationality, as we are using the term, are broadly distinct from the sorts of issues that have been interestingly explored under the label "epistemic utility theory" (EUT). See, e.g., Carr (2017) and Pettigrew (2021). As we understand it, EUT is about applying decision theory to the apportionment of one's credences in order to maximize some sort of accuracy among them. We are concerned, rather, with more literal apportionment of resources, such as time, money, and person-power, at the community level. It would not surprise us at all if results from EUT turned out to be useful to our purposes, but for now we just want to note that these are two distinct toolboxes, and neither should count as automatically inheriting the advantages or disadvantages of the other. Second, we are officially agnostic on the relationship between the epistemic and the zetetic, and in specific contrast to recent work by Jane Friedman, especially Friedman (2020), we don't think the positions we will be promoting here are in any tension with traditional epistemic norms. It is, rather, orthogonal to it. We are, that is, constructing our own piece of "applied zeteololgy" from concepts distinct from those of traditional epistemological theorizing. If our arguments nonetheless still count as

2 A Better Normative Framework

In order to answer these questions, let's start by defining methodological rationality in terms that we hope are nearly axiomatic:

> MR: A research community ought to pursue its inquiries in ways that maximize benefits while minimizing costs.

Of course, MR can get away with being nearly axiomatic by also being nearly vacuous. So, in order to put this skeletal idea to work, we need to put some meat on its bones about what "research communities" are, and what should count as the relevant sorts of benefits and costs.

We'll start with what it means to be a research community, and before we do that, it is important to say something about why we want to pitch things in terms of research communities in the first place. We want to focus on research communities because a significant part of the shift from talking about epistemic normativity to talking about methodological rationality involves shifting our attention away from personal-level normative considerations and onto group-level normative considerations, with corresponding shifts in temporal frame as well. This does not mean that individuals are unable to conduct research on their own. People can, and sometimes, "do their own research," although perhaps they often shouldn't, for a whole host of reasons.[9] Even so, a solitary castaway can in principle function as a sort of singleton "community" in the sense that we have in mind here. Nor does it mean that group-level normative considerations don't entail, or perhaps even reduce to, individual-level normative considerations. After all, groups can only do group-level things, including responding to various normative considerations that are imposed on those groups, by means of the individual members who make up those groups doing appropriate individual-level things. Individual researchers must decide what to do in order to help satisfy community-level normative considerations, how to resolve apparently conflicting community-level normative considerations, how to apply community-level normative considerations that are unclear in particular circumstances, and so on. They also have to decide when to allow other kinds of considerations to override the community-level normative considerations that are imposed on them as members of research communities. What it does mean, and our primary reason for focusing on research communities rather than individual researchers, is that attending solely, or even primarily, to individual-level normative conditions, even those normative considerations

epistemic in some sufficiently broad sense of the term, that's certainly fine by us. We still definitely consider ourselves card-carrying epistemologists, after all.

[9] For an interesting and timely discussion, see Ballantyne et al. (2024) and Matheson (2024).

that apply to individual members of research communities *qua* their membership in those communities, leaves crucial aspects of methodological rationality out of focus. In particular, we worry that it makes it all too easy to overlook the fact, as we will argue in what follows, that the kinds of normative considerations that are the center of recent debates about the method of cases are community-level normative considerations. This is why we think that the focus going forward should be on research communities and the normative considerations that apply to research communities.

We assume that there is nothing spooky here about talking about community-level obligations, or group obligations more generally. For that matter, we think that there are many ordinary examples of these kinds of obligations. Consider, for example, the obligation of an academic department to create a healthy gender climate. Of course, individual members will have their own, non-derivative obligations to treat their colleagues well and to intervene when others are not doing so. But the department, considered as a whole, also has responsibilities and it may uphold them through a mix of formal and informal mechanisms, perhaps including, but importantly not limited to, the creation of institutionalized offices and positions, such as that of departmental ombudsperson, or members of a climate committee. Most individuals will have a responsibility to do their part in creating such a healthy climate, but establishing that climate is simply not within the powers of the individuals *qua* individuals. The community, as a whole, is responsible for the climate that it, collectively, establishes or fails to establish. Research communities will also be responsible, collectively, for the appropriate allocation of a host of resources related to the research that they are conducting, including the allocation of money, time, and status, the maintenance of forums for communication, the enforcement of academic integrity, the training of the next generation of researchers, and so on. In terms of methodological rationality, such resources are the primary currency in which costs are to be counted. Any research community has some budget of such means to expend, as best they can, for the purposes of inquiry. Different types of distributional decisions will be made at different scales of organization, and some are optional, and others will be obligatory. For example, an individual research laboratory may decide whether or not to undertake to replicate another laboratory's work as part of their contribution to a community-level norm of replication. But a lab must expend appropriate levels of resources to make sure that data is reported correctly, that records are maintained, and so on. And, while it is tempting to suggest that these responsibilities fall on a specific member of the research community, namely, the principal investigator, it is important to note that PIs can only fulfill these kinds of obligations by establishing well-run laboratories whose staff can do the actual work required by such norms. And other decisions and obligations are executed at larger scales. Departments may decide what their teaching needs are for training the next generation of researchers, and what research areas they would like to prioritize in their hiring decisions. And it is

often whole disciplines that maintain and govern professional organizations and journals.

All these examples have been fairly abstract, and apply to a great and diverse panoply of different research communities at different scales. But specific academic disciplines, departments, laboratories, and even pairs of coauthors may have more specific and contentful obligations. For example, different journals will, depending on their scope and mission, be required to maintain stables of editors and referees with specific sorts of skills and knowledge bases. The *Journal of Philosophical Logic* has to have access to experts who can check highly sophisticated proofs; *Cognition*, on the other hand, needs referees competent in experimental design and quantitative analysis; *Kantstudien* makes a strenuous demand on its referees for the polyglot mastery of the scholarship of the history of transcendental idealism; and so on. So, pools of resources will be managed by different entities at different levels, and inasmuch as we are thinking of them *qua* research communities, we can ask whether their constitutive agents and institutions are deploying those resources as rationally as they ought. There are of course other normative demands on these entities, including ethical ones, for example, how we should treat our research subjects and, for that matter, our graduate students. There are also practical concerns that we are just not going to consider here, such as the many that arise at the interface of research, education, and, especially for state-funded institutions, politics. Not that there are not interesting and indeed pressing questions in their own right here, but they don't seem to us ones that will be especially salient to the methodological issues about experimental philosophy and armchair philosophizing.[10] As members of a research community, we thus face the mandatory question: how best can we expend the particular resources allotted to us, in order to advance the goals of inquiry? Are we getting the best veritistic bang for our methodological buck?

With all of this in mind, we will use "research community" in what follows as a term of art to refer to a group of people whose interactions with one another are organized around the goal of acquiring novel truths about some subject matter or another.[11] There are numerous ways that such a community might be organized. Perhaps the people are charged with this goal as part of a specific organizational mandate. Perhaps this goal is simply part of their collective understanding of the nature and purpose of the community. Perhaps this goal is not at all explicit, but is implicitly reflected in some set of shared practices and norms that the members of

[10] For important discussions about the relationship between science and other social and political institutions, see Longino (1990), Kitcher (2001), Douglas (2009), and Harding (2015).

[11] An anonymous referee encouraged us to qualify this somewhat by adding that the truths should have some "utility or substance." And we can see this kind of sentiment expressed by, among others, Sally Haslanger (2000) and Daniel Dennett (2006). While there are surely merits to the cases that they make, we don't wish here to exclude communities that are organized around esoteric or niche topics, such as sabermetrics or the finer points of philosophical methodology.

the community have adopted. In short, how the members of the community come together around this shared goal is not particularly important.[12] That a research community is organized around a shared veritistic goal doesn't have to mean that all its particular members place much explicit emphasis on this goal. All that is needed is that we can rationally reconstruct the community and its members in terms of this shared goal, and that we can think about the normative considerations in terms of this goal.

A few points of clarification are quickly in order. First, we include talk of how the community is "organized around the goal" of inquiry here since arguably almost all communities have at least *some* degree of need for inquiry, if perhaps of a fairly mundane sort. For example, a chess club will be committed to learning truths about who has won which recent tournaments and who is responsible for bringing the bagels to the next meeting. We take it that these particular epistemic goals are subservient to, but not among, the primary goals of that community as such. Thus they are not goals that play an organizing role for that club, which instead is likely organized around goals such as promoting chess, facilitating the enjoyable play of that game by its members, and the like.

Second, the acquisition of novel truths only needs to count as *one of the* central, organizing goals in order for a community to count as a research community in the sense that we have in mind here. This is meant to allow room for philosophers who would define the goal of philosophical inquiry in more demanding terms, such as the acquisition of knowledge (for example, Kelp 2021a and Kelp 2021b), so long as the pursuit of truth is among them. For that matter, we also leave room for a wide and pluralistic range of non-alethic goals as well, so long as the alethic goal is among the most central organizing ones.[13]

Third, research communities, understood in this way, exist at a range of different scales from global academic disciplines to local research laboratories, and perhaps even to pairs of research partners and coauthors. We also mean to include not just collaborative interactions but also adversarial ones: when one philosopher debates another in the literature, they are doing so as members of the same research community. So too are different scientists competing for the same set of grants.

And, finally, we aren't going to concern ourselves overmuch here as to the exact metaphysics of groups involved, nor will we be concerned if there are some occasional odd sorts of collective entities that satisfy our definition but don't (ahem) intuitively seem particularly research-oriented, so long as we avoid such nullifying results as *all* groups vacuously counting as research communities. The notion of being "organized around" should similarly not be read as having any heavy-duty

[12] Our thanks to Aaron Meskin, Ron Mallon, and Allan Hazlett for some very useful interrogation on this point.

[13] This is perhaps a point of both concurrence with and divergence from Thorstad (forthcoming)'s global consequentialism about norms of inquiry.

metaphysical or teleological commitments. What is important for us is what can be worked out from this starting point of a community's having a shared goal of truth-acquisition.

With all of that in mind, we hope that not much really needs to be said here in defense of the idea that philosophers, and in particular the subcommunity of analytic philosophers, is in this sense a research community. There is nothing special about philosophy in this regard; surely many, perhaps even most, academic communities can be uncontroversially characterized as research communities. We have not conducted a survey of the profession, which is perhaps a bit embarrassing for a pair of experimental philosophers to admit. And for that reason, we are certainly prepared to have this claim empirically falsified. But we strongly suspect that a substantial majority of philosophers would happily accept the pursuit of novel truths to be at least one of the organizing goals of philosophy. With such caveats, though, we aren't just fishing in the dark here either. In fact, we think that there are actually good reasons for thinking about philosophy as a research community, namely, the widely endorsed and enforced practical norms that govern how we do philosophy. Perhaps the best example comes from thinking about the norms that govern the publication of philosophical papers. When we are asked to evaluate the quality of submitted papers, we start by looking for ways in which the arguments rehearsed and advanced in the paper might fail to get at the truth. Is the reasoning invalid? Does it commit any fallacies? How questionable are the premises? Are there any crucial errors in scholarship? Novelty is also placed at a premium here. An argument that is otherwise without mistakes, but which is basically a rehash of someone's already published results, even the author's own previously published work, will often fail to get past the gatekeeping process, and for good reason—republishing an argument already in the journals would likely be a poor use of a scarce resource. We also give nontrivial importance to citational lapses when evaluating submitted papers, and papers are routinely rejected for failing to pay enough attention to the current state of play. The importance of good citational norms, and the enforcement of these norms, can be understood in terms of the value that philosophy places on the pursuit of novel truths. First, we want authors to point us to their sources so that we can see and judge for ourselves what their arguments are based on and whether they pass veritistic muster. Second, we want authors to give credit to those who got there first.[14]

If all of this is right, and we think that it is, then this is excellent evidence that philosophers make up, in our sense, a research community, and that MR is thus normatively binding on them. Given that we have a collective goal of acquiring novel philosophical truths over the long run, it can be sensibly asked whether we are making the best use of what resources we have available to us in pursuing that

[14] Michael Strevens (2003) calls this the *priority rule* in science.

goal. In what follows, as was foreshadowed by our discussion of unmitigated vulnerability above, we will be especially interested in questions of methodological rationality that pertain to managing the risk of errors. The benefits are to be scored against that veritistic goal; the costs, however, are paid for in numerous denominations, such as human time and effort; paper and ink; attentional bandwidth, including precious journal space; and, for that matter, such denominations as $, €, or ¥.

Journal practices again provide a nice example here, in this case, of different resources spent toward error-mitigation. Since all researchers are all too human, mistakes do get made, and an integral part of the job for editors and referees is to help prevent them from getting into print. Sometimes this means rejecting a manuscript, but often it involves just pointing out a potential error and suggesting how the authors might correct for it, say, by running another study that controls for a possible confound. When necessary, errata or even retractions are printed. Our community finds it obviously appropriate to spend the resources of editorial time and print space toward this end. And best of all, a journal will often provide a forum for dissenting researchers to publish their own findings and arguments to challenge what has already made it into print.

In terms of scoring benefits against this goal of inquiry, we will use a completely extensional read on both truths on the one hand and errors on the other. In a nutshell, what we mean are *inaccuracies*. Methods are shot through with claims that purport to veridically represent on how the world is, ranging from the individual datum reporting an instrument's reading on the value of some observable parameter at a place and time, all the way up to grand sweeping theoretical generalizations. We will construe errors here to be instances when a method asserts *p*, but as a matter of fact, *p* is not true.[15] There are other legitimate notions of error, most importantly ones involving not alethic accuracy but *compliance with rules*, for example, when some member of a research community fails to follow the prescribed rules of a method, whether or not it results in a falsehood or, instead, an accidentally acquired truth. We aren't at all objecting to the cogency of this latter way of thinking about error, and in later chapters we will have a lot more to say about instituting and following good rules in philosophical practice. But it is just not the one we are going to use to frame our application of MR to the armchair method of cases. And, of course, these are not unrelated: presumably a chief reason a community might seek to enforce such rules of reasoning would be in service of acquiring truths and avoiding errors.

[15] What we most have in mind here are cases where *p* is false, but it is fine also to include cases where *p* lacks a truth value or where, out of linguistic or conceptual confusion, there turns out not even really to have been a proposition *p* expressed. Having said that, we will simplify our discussions here by setting these kinds of cases aside.

Understood in this way, errors are methodological costs in and of themselves, not in terms of a direct expenditure of resources but rather in terms of the foregone benefit of the truths they are mistaken about. Crucially, the cost of an error may not stop with itself, but in addition, and often much more consequentially, in the risk of greater costs to inquiry as they expand to contaminate whole lines of research. Yesterday's bad conclusion is today's bad premise, which we may well stumble over tomorrow and land in yet further falsehoods. A theory established on false evidence may squat upon the course of future research and redirect it along wayward paths. And at the same time, an error robs us of the inferential value of the truth it has displaced, that might have itself served as a premise in many a sound argument for further true results.

In case you are worried that we are attempting to impose upon philosophy an alien, scientific way of thinking about progress and error, we happily note that Williamson (2006) makes fundamentally the same point. He is primarily concerned in that paper with the costs associated with giving inadequate care to the formal aspects of philosophical inquiry, but the worries he raises generalize to other sources of error:

> How can we do better? We can make a useful start by getting the simple things right. Much even of analytic philosophy moves too fast in its haste to reach the sexy bits. Details are not given the care they deserve: crucial claims are vaguely stated; significantly different formulations are treated as though they were equivalent; examples are under-described; arguments are gestured at rather than properly made; their form is left unexplained; and so on. *A few resultant errors easily multiply to send inquiry in completely the wrong direction* . . .
>
> Precision is often regarded as a hyper-cautious characteristic. It is importantly the opposite. Vague statements are the hardest to convict of error. Obscurity is the oracle's self-defense. To be precise is to make it as easy as possible for others to prove one wrong. That is what requires courage. *But the community can lower the cost of precision by keeping in mind that precise errors often do more than vague truths for scientific progress.* (2006, 289; emphases added)

The two italicized portions of this passage provide a nice way of understanding what we think goes wrong with armchair philosophy, and the armchair-bound method of cases, more specifically. Uncorrected errors pose serious downstream threats to the methods used by any research community because they can contaminate all the channels of inquiry they flow into. This makes detecting and correcting these errors, or otherwise mitigating their threat of contamination, a fundamental engine of methodological progress and merits spending substantial methodological resources to achieve. And, as we will argue in the next chapter, armchair philosophical methods lack the resources needed to do this important work.

Philosophers who, like us, are fond of William James's (1896) "Will to Believe" may be worried that we are placing ourselves too much on team "Shun error!" when the MR framework suggests we should be wholeheartedly on team "Believe truth!" Yet we do number ourselves among the Jamesian true-believers. Translated into our terms here, it is clear that James is scolding his target, William Clifford, for being too risk-averse, indeed pathologically so when he writes colorfully that Clifford is in "his own preponderant private horror of becoming a dupe." But once it is made clear that "nothing ventured, nothing gained" applies to methodology as much as to other of life's endeavors, it is clear that we are also rationally obligated to treat our methodological ventures in accord with principles like MR. And the better we are at managing our exposure to veritistic risks, the more adventurous we can afford to be in our investigations. We want, as researchers, to be able to take chances, even big ones. We just want to minimize the damage that can be accrued to our projects when, as will inevitably happen, some of these chances pan out badly. This is part of why we want to speak of *mitigating* the risk of error, and not just *reducing* it. Managing the risk in a retirement portfolio doesn't usually mean just making only low-risk investments; sometimes it means including high-risk-high-reward ones, and then also making appropriate hedges, so a bad outcome doesn't crack your whole nest egg.[16]

Here's an instructive analogy, both in where it works and where, ultimately, it does not. Errors are to inquiry a bit like shoplifting is to storekeeping. They are a necessary cost of doing business, which we can strive to minimize, but cannot legitimately hope to prevent altogether. In fact, if shopkeepers act too aggressively to prevent the costs of such theft, perhaps by putting customers through an extensive search when they depart the store, they will surely lose much more in terms of lost sales, and in terms of customer satisfaction and loyalty, than they will gain in terms of prevented loss. The same is true for research communities. If we act too aggressively to prevent error, perhaps by holding all of our inferences to a standard of metaphysical certainty, we will lose much more in terms of novel prediction and increased understanding than we will gain in terms of prevented mistakes of reasoning. Likewise, shopkeepers who allow customers to enter their stores, conduct their business, and leave with minimal effort will incur some loss, but will also make more sales as well. And the same is again true for research communities. Setting standards for acceptable inference somewhere below the standards set by

[16] There's this bit at the end of the relevant paragraph that is worth addressing: "Our errors are surely not such awfully solemn things. In a world where we are so certain to incur them in spite of all our caution, a certain lightness of heart seems healthier than this excessive nervousness on their behalf. At any rate, it seems the fittest thing for the empiricist philosopher." We are inclined to agree or at least to be sympathetic, but we think having good error mitigation efforts where they can be afforded will generally be a small price to pay to achieve that "lightness of heart." Also, James is talking about an individual believer, and we are focused on both the longer and broader term "uncorrected errors," which will matter rather more for us, along the Williamsonian lines just discussed.

Descartes in his *Meditations* will surely lead to some false beliefs, but also in all likelihood to a great many more true ones as well.

This analogy, like all analogies, only goes so far, and here is one key disanalogy, which makes the situation for error-prone research communities potentially better than that of theft-plagued shopkeepers. Once a pair of stolen shoes has been walked out the door, they are most likely gone for good. But facts aren't like that. Shoplifted goods generally can only be written off, but almost never recovered, whereas cognitive mistakes very frequently can be corrected. So the practical options available to research communities and shopkeepers importantly diverge in kind here. The best that shopkeepers can do is adopt measures that find the profit-maximizing trade-off between unfortunately inconvenienced, honest customers and appropriately inconvenienced, dishonest ones, and then grit their teeth and accept the losses as they come. Tracking down shoplifters well after the fact will not generally be worth the expense and effort. But researchers, in contrast, often can reasonably attempt to hunt down mistakes even long after they are made, and they can with much greater than zero frequency bring them to light and correct them. And when they *can* try to do so at a reasonable cost, methodological rationality will generally prescribe that they *should* try to do so.

Another comparison and contrast with shoplifting heightens the stakes further, when we consider more fully how both shopkeepers and research communities are looking to succeed in the long term. A precondition for long-term success of any enterprise is avoiding *ruin*, and smart shopkeepers will need to be sure to price the catastrophic costs of ruin into any estimations of risk and reward. In terms of managing a business, it's clear enough what counts as avoiding ruin, but what does "ruin" mean when we are in the business of inquiry? It seems to us that inquisitive ruin most often takes one of two forms: *terminal incorrigibility* or *immobilized stalemate*.[17] In the former, mistakes get made that we lack the resources to un-make, and thereby get locked in indefinitely. Inquiries can generally tolerate even a large number of these kinds of errors so long as they are trivial, as well as unsystematic and uncorrelated, especially at the level of individual observations. But the greater the ramifications of these kinds of errors, and of the aggregated—and, *ex hypothesi*, unexorcisable—body of such errors, the more it will obstruct the future course of discovery. In the latter sort of methodological ruin, we find ourselves permanently lacking the means to make good decisions among a set of competing hypotheses, or perhaps even lacking the means to frame novel hypotheses in the first place. While the ruin of incorrigibility will tend to look like an irreversible slide into an errant but impregnable dogma, the paradigm of stalemate manifests

[17] We are open to others, but they may be beside the point of our discussions here. Say, if the inquiry gets shut down completely, perhaps due to civilizational collapse or the like, that's an obvious form of ruin. But under such circumstances questions of methodological rationality would be rendered profoundly, tragically otiose. We will discuss philosophical stalemates, and how experimental philosophy can help us over them, in Chapter 6.

as an interminable squabbling between entrenched but mutually unconquerable foes. (Perhaps this will sound familiar to contemporary philosophers.) In both sorts of cases, we can see why uncovering errors as such will be so crucial, either to unsettle any would-be dogma and render it vulnerable to further investigation, or to allow some contesting hypotheses finally to be vanquished and surrender the field. The moral of the story is that ruin in inquiry can only be avoided if we have adequate resources to find out when and where we've gone wrong.

Some philosophers, particularly those with a taste for formal methods, will probably want to object at this point that not all philosophical inquiry must treat error as inevitable. The practices of proof in mathematics and logic, especially as they have been refined over the centuries, not only hold, but have delivered on, the legitimate promise of a vast, substantial, and verifiably error-free body of results. We too happily celebrate the tremendous success of such practices, while nonetheless cautioning against overgeneralizing from them. In general, human inquiry has developed two main families of strategies for managing the threat of error, which we will call *P-strategies* and *S-strategies*.[18] *P*-strategies aim to create a protected body of results that have been purged of any pieces that we are less than certain about. Because the highly curated body of results is practically perfect, we can treat them as utterly trustworthy, and a great many methodological benefits come with that. For example, the set of protected results will increase very nearly monotonically, and we can rely on multi-premise closure without concern for counterexamples, which accordingly means that we can deploy even extravagantly long derivations without fear of aggregating small chances of error.[19] This does depend, of course, on having extremely effective practices for verifying such derivations in the first place. But it seems that we do indeed have such practices.

Yet *P*-strategic glories do not come without a price. They require us to purify our methods until we are relying only on the hardest, sharpest components of our cognition. We truly can achieve a working infallibility, but only within the stark boundaries of a near-ascetic methodological discipline. So many other of our tools simply cannot withstand such purification, including both casual and scientific observation and most forms of ampliative inference. For *P*-strategies to work, their end products must be practically risk-free and, yet, nothing inductively ventured, nothing inductively gained. That's why we have the second family of approaches to error-management. *S*-strategies permit a radically enlarged set of methods, at the cost of abandoning any hope of constructing a safe zone in which all error has been shut out. While errors will still be pre-emptively avoided as much as possible, *S*-strategic practitioners know that that possibility runs aground far before achieving

[18] This distinction, and much of this section, is borrowed from Weinberg (2015).

[19] For further discussion on the issue of accumulating risk across inferences, see Lasonen-Aario (2008). Her arguments also, we would note, illustrate how hard it is to fix problems of error risk just by increasing up-front baseline reliability.

anything like a substantial body of validated theorems. They must, therefore, devote substantial resources to assessing where errors could arise, checking for them after the fact, and eliminating, or at least quarantining, them. If precise proof is the paradigm for a *P*-strategic method, then self-scrutinizing scientific practices are the paradigm for *S*-strategies.

It is not just that *S*-strategies can be appropriately more permissive about what sources of evidence to deploy, compared to *P*-strategies. They also can make heavy use of ampliative modes of inference, such as inference to the best explanation and reflective equilibrium. It is in the nature of ampliative modes of inference that, even used conservatively, they introduce non-trivial risks of error.[20] In fact, they introduce the possibility of whole new *kinds* of error, especially regarding various judgment calls that human practitioners unavoidably have to make when making such inferences. For example, we might have to ask: does our sample include enough observations, or do we need to acquire more? Stop too soon, and we may be making a hasty generalization, or simply averting our gaze before learning an important truth. But allow ourselves too much leeway about pursuing more data points and we may be engaged in the "questionable research practice" known as "optional stopping"—a well-known recipe for producing unreplicatable results.[21] And how are we making sure our sample is sufficiently random or otherwise representative of the population we are looking to generalize about? Another example: we can cite and deploy only a miniscule fraction of the available *explananda* when advancing our preferred candidate for *explanans optimum*, so how do we keep our own motivated cognition from leading us into the sin of cherry-picking? It's important to understand that, while we can take many measures to try to minimize such errors on the front end, they will always come up at least somewhat short.

And that's all while imagining that these methods are being pursued rather conservatively—yet, in the right overall methodological context, what can look like wanton jumping to conclusions can in fact be a sign of a healthy research community at work. To take a highly salient example, inquiry at the level of the philosophical community is organized as a vigorous, even rambunctious, "marketplace of ideas," with very low barriers to entry for any views that are not self-contradictory (and maybe even some that are, if they have other attractive features). So the

[20] For discussion of "inductive risk" see Rudner (1953), Jeffrey (1956), Levi (1962), and especially Hempel (1965), who introduced the term. For more recent discussion, especially on the relationship between inductive risk and values in science, see Douglas (2000, 2009), Elliott and Richards (2017), and Elliott and Steel (2017). Part of what we are trying to do in this book is make philosophers aware of the kinds of problems that come up in philosophical research and how we can better work to mitigate these problems. This requires us to make it clear that philosophers employ both *S*-strategic methods and *P*-strategic ones, and to think carefully about what it means to use *S*-strategic methods responsibly. This means, we think, that philosophers need to be more aware of things like "inductive risk" and reasons to think that the risks associated with using *S*-strategic methods increase the role that both epistemic and non-epistemic values play in philosophical practice (on analogy to the way that these methods increase the role that they play in scientific practice).

[21] See Sanborn and Hills (2014), Rouder (2014), and de Heide and Grünwald (2021).

mechanism of theory development here is one of launching dozens of hypotheses into logical space and trying to arrange for most of the wrong ones to crash. That this only works when the wrong ones really do crash, and can moreover be *seen* to have crashed, is at the heart of Williamson's observation above about how the wheels of progress are lubricated by the *clarity* of our errors. This is even more dependent on error detection than when we use moderately reliable sources (such as perception) and perform ampliative inference on the outputs they yield, for in those cases we at least have a presumption of substantial baseline reliability for both the sources and the inferences. Considering the entire fractally diverging set of mutually inconsistent philosophical hypotheses that get put forward and strenuously debated by our community, obviously the vast majority of them will have to turn out to be wrong.

So, while *P*-strategies for error management are a very real and crucial part of inquiry on the whole, they are not going to be viable beyond a splendid, if nonetheless tightly circumscribed, arena. Everywhere else in philosophical inquiry (and beyond), we have to adopt more complex arrays of tactics, preventing errors where we can but knowing that we must—*must*—also prepare for how to handle their inevitable presence in our ongoing efforts. When the price is right, we can and should improve front-end reliability where we can, as Nado suggests, but so long as we are still working with moderately reliable sources, then there will be a real risk of error, and that risk will need to be managed, often on the back end. As with any *S*-strategic form of inquiry, an absolutely crucial component of philosophical methodology in general will be assessing our risks of errors, continually monitoring for them, developing means for correcting for or quarantining their threat, and generally managing our risk of vulnerability.

There is of course no one activity that is "monitoring and correcting for errors," or even "monitoring and correcting for errors in analytic philosophy." It depends enormously on what our state of knowledge is about what kinds of errors our methods and practices may be prone to and what resources are available to handle them. No tactic works for all cases, and for that matter, all tactics come with their own associated costs and shortcomings. For example, sometimes formalizing an inference is helpful; other times it is not. And, indeed, formalization can bring in novel sources of error, such as when some important nuance in the natural language argument goes missing in the formalized version. Sometimes consulting with a relevant empirical science can help to avoid certain sorts of blunders; other times the scientific work itself is shot through with the very philosophical confusion we were looking to address. Sometimes a subtly different phrasing of a thought experiment can bring about an insight that was obscured in other formulations; other times such a change merely opens the door for the unwanted influence of irrelevant psychological factors. Much of what it is to really have a *method* in some area is to be able to distinguish between the "sometimes" on the left side of those semicolons from the "other times" on the right, and to systematically steer

your efforts leftward. The fine-grained details of many methodologies standardly include a nosology of possible errors, and highly concrete instructions as to how to avoid them, and under what conditions.

Moreover, the whole process is, like inquiry itself, highly dynamic. New methods emerge, bringing with them new resources for resolving errors that may afflict other methods—and their own peculiar proclivities for error. We have a constantly evolving understanding of existing threats of error and our available means for addressing them; let's call this the *error profile* of a method. Such error profiles will depend on fine-grained empirical and practical knowledge, which do not come for free. We must always be willing to ask afresh of a practice of inquiry, as a matter of methodological rationality, are we doing enough to cultivate the necessary resources to prevent, discover, and/or correct for errors, and to shield our theories and inferences from the errors that cannot be thus rooted out?

We hope that all of this strikes the ear as at least near-platitudinous. Yes, someone might say, of course a properly conducted research will take steps to avoid error and to root out errors that were not initially prevented. Nevertheless, it is one thing to acknowledge a platitude verbally; it is rather another to take concrete steps to implement it. And it seems to us that it is a platitude not well implemented in much of the current debate over philosophical methodology, which as we noted in the previous chapter has by and large focused much more on questions of reliability. And because of this, an important lesson has not been learned: from the point of view of methodological rationality, the mitigation of error risk is not merely one epistemic good among others, but an *essential* good without which the long-term goals of inquiry cannot be successfully pursued.

Moreover, if we are using *S*-strategic methods, then we cannot simply focus on avoidance, that is, on improving front-end reliability. It is imperative to learn how to live successfully with error, including detection and correction after the fact as well as other strategies, such as developing modes of inference that can withstand even substantial error in one's set of premises. As we hope this discussion helps to illustrate, error management is not a kind of methodological bonus, a sort of supererogatory good for inquiry. Viewed within the framework of methodological rationality, we can see that considerations of error management must be central to our deliberations as to how best to apportion our methodological resources, and not a mere afterthought. When a research community fails to weigh the value of error management adequately, they are highly likely to find themselves with a lopsided portfolio of methodological investments. They would be, in short, *methodologically irrational*, collectively distributing their community's resources in ways that fail to maximize their likely long-term acquisition of novel truths.[22]

[22] To be clear, methodological irrationality can arise for any misapportionment of resources of inquiry, not just an underinvestment in error management. It is surely possible at least in principle to over-invest in error management as well, for example. Or, a community may put too many resources in a less-reliable method over a more-reliable one, without some other sorts of compensatory values to

Recall the episodes from the history of science surveyed above, in which researchers felt and acted on a sense of urgency about newly discovered possible sources of error in their methods. That sense that something really *ought to be done* could not be captured easily in traditional epistemic terms because the threats of error were not so profound as to rise to the level of full-blown severe epistemic deficiency. But such urgency becomes normatively intelligible when viewed in terms of methodological rationality: to find a not-yet-mitigated threat of error in our methods is to discover that our current set of methodological investments are not as high in quality as we had previously thought, and at the same time to reveal what could be a very high-yield direction for investment, in addressing that threat. Because uncontrolled error threatens a dire reduction to our long-term veritistic payoff, it will very often be truth-maximizing to redirect resources, perhaps substantially, in order to bring any such threatened method out of a state of unmitigated vulnerability. Accordingly, it would be methodologically irrational to fail to redirect resources in such a manner.

One cannot make this generalization in an unhedged form, however, because there are possible circumstances that would still render suboptimal any such proposed reallocation. It will depend on the particulars of the trade-off involved; the goods of a novel investment in error mitigation can only be had at some cost in terms of the goods of inquiry that are lost in the transfer of resources. (Obviously, if this were not the case—if we could conjure new error-management resources out of thin air—then we should simply do so!) For example, if a newly discovered threat of error turns out already to be reasonably well mitigated by procedures already in place, then the decision may be less obvious, and rather more of a judgment call as to just what further measures might or might not be worth adopting. But when a community learns that it is exposed to a hitherto unrecognized source of error, and hence likely to be also a yet-unmanaged one, then the payoffs for such investments are likely to be quite stark—so stark that it would be methodologically irrational not to pursue them. And hence, the urgency.

3 Methodological Rationality and Armchair Philosophy

With all of this in mind, let's return to the way that we characterized the armchair method of cases in Chapter 1, as a method that philosophers can pursue

the former; for example, the community may expect thereby to learn how to better use the less-reliable method and thus make it more reliable over the long run. Nonetheless, we do suspect that underinvestment in error management is a generally common form of methodological irrationality, not least of all because it's rarely as much fun as just barreling forward with one's first-order methods—the thrill of discovery will trump the meh of careful bookkeeping more often than not. And anyhow, underinvestment in error management is the source of methodological irrationality at stake when considering the armchair method of cases.

without any special sorts of empirical inquiry; basically, ordinary observation plus all of the logic you want, perhaps together with some set of accepted common sense and scientific background knowledge. As we noted there, we take this to be the kind of practice that Williamson (2007) has in mind when he writes

> Every armchair pursuit raises the question of whether its methods are adequate to its aims. The traditional methods of philosophy are armchair ones: they consist of thinking, without any special interaction with the world beyond the chair, such as measurement, observation or experiment would involve. (p. 1)

Changing the terms of the debate from the conceptual framework of epistemic rationality to the conceptual framework of methodological rationality lets us reconfigure precisely the kind of question that Williamson says is raised by our armchair practices. To ask whether or not armchair philosophy's resources are adequate for its goals, we don't need to ask anything like: what might be so wrong with the method of cases that we shouldn't use it at all? Instead, we can ask: what methodological benefits would we get from leaving the armchair, and would those benefits outweigh whatever costs come along with doing so? And we can frame the debate over armchair philosophy as exactly the kind of cost-benefit analysis that the concept of methodological rationality is made for.

Having established the framework that we would like to use for thinking about the relationship between experimental philosophy and armchair philosophy, we plan to show in the next chapter how armchair philosophy is *methodologically irrational*. In short, the current practices of armchair philosophy are severely limited in their capacity to shield the method of cases from errors, such that the community would be wise to make a substantial—and empirical—investment in upgrading those capacities. In Chapter 1, we argued that we have a fairly substantial, and consistently growing, amount of empirical evidence that the case verdicts that philosophers commonly appeal to are susceptible to a range of different errors. In the next chapter, we will argue that armchair philosophy does not have the conceptual or methodological resources to handle them. Subtle order or context effects will be indetectable to introspection or unsystematic observation. Demographic differences will often be obscured both by selection effects (only those with the 'right' verdicts will advance in any given areas), and by the plain fact of professional philosophy's lack of diversity. The particular demographic variation between philosophers and non-philosophers will by its nature be especially hard for philosophers to detect from their own armchairs. Moreover, even if philosophers came to be aware of such threats of error, there do not seem to be many armchair-friendly methodological adjustments that could help. On our account, the limits of the armchair method

of appealing to case verdicts do not so much inhere in anything pathological about the philosophical case verdicts themselves. Rather, the collective value of those verdicts is sharply curtailed by the limitations of the armchair's folk psychological resources in managing the empirically live threats of error in those verdicts.

5
The Limits of Armchair Philosophy

We have now established the framework that we think should be used when philosophers think about experimental philosophy and the method of cases. In this chapter, we will bring together elements that we have discussed in the previous four chapters to argue that armchair philosophy fails to satisfy the basic principles of methodological rationality. We could, and therefore should, redistribute our investigatory resources in order to develop resources for managing the kind of risks of errors that we are currently highly vulnerable to.

This sort of situation is not unique to philosophy, and whenever new instruments or methods are introduced in the sciences, a similar obligation arises to enable the relevant research community to use them in a way where it can adequately manage whatever error possibilities that the new tools will have brought into play. In order to satisfy the principles of methodological rationality, members of a research community, especially ones making heavy use of *S*-strategic methods, will be missing out on key long-term benefits if they have inadequate *error profiles* for the instruments and methods that they use to conduct their research. These error profiles provide an account not just of the baseline reliability, but much more importantly they characterize the workings of such instruments and methods and provide practical information about the particular sorts of errors that they are prone to and how those errors can be avoided or resolved. So, for example, the error profile of a compass not only tells us the basics of how they point toward the north, and perhaps what percentage of the time they are likely to do so, but also about their sensitivity to magnetic fields, tilt measurement, and magnetic inclination. And the error profile of mass spectrometry not only tells us how the method works and what percentage of the time it is likely to give us the correct mass-to-charge ratio of the molecules in our sample, but also about the dangers of fragmentation that can occur when unstable molecular atoms dissociate as they pass through the ionization chamber, and about the ways to prepare samples that are sensitive to light, temperature, or pH. The same sorts of stories can be offered for any instrument well entrenched in scientific usage. And for instruments looking to become well entrenched, its community will be found to be in the process of developing one, or discovering that the instrument should just not be used.

But, as we shall argue here, our philosophical community does not yet provide us with an adequately developed error profile for the armchair method of cases. At the end of the last chapter we rehearsed some general reasons to think that armchair philosophy cannot handle such errors, because armchair psychology cannot

Standing Up for Philosophy. Jonathan M. Weinberg and Joshua Alexander, Oxford University Press.
 DOI: 10.1093/9780191947704.003.0005

do so. Beyond the highly circumscribed limits of armchair psychology, however, the question might nonetheless be raised whether other, philosophy-specific aspects of our practices might nonetheless provide the requisite sort of working knowledge about how to handle the risk of error in armchair philosophy. In the first half of this chapter, we will look for any such raw materials in philosophical practice out of which to build an error profile, and find the cupboards to be pretty bare. We will start by looking at what we will call the *practical methodological wisdom* about the method of cases, and argue that the kind of error profile that emerges falls far short of enabling armchair practitioners to anticipate and manage the kinds of problems that have been discovered by experimental philosophers. We will then turn our attention to what we will call the *folk theory of philosophical expertise*, which contends that, while armchair philosophy might lack the kind of error profile needed to anticipate and detect the kinds of problems that have been discovered by experimental philosophers, philosophical training provides us with the resources needed to correct them. Here, we will argue that the folk theory of philosophical expertise rests on unsupported, and in some cases outright false, empirical hypotheses and does not go any real distance toward helping show that armchair philosophy can solve its own problems. Finally, we will argue that the only way to bring analytic philosophy back in line with the requirements of methodological rationality is to incorporate the experimental methods that are used by experimental philosophers.

1 The (Currently) Sparse Error Profile of the Method of Cases

As we briefly discussed in Chapter 4, when considering (and rejecting) *hopelessness* as a possible basis for an impeachment argument, it is not that philosophers have *no* good ideas about ways case verdicts can go wrong, and how to correct for them. While these ideas have not, so far as we can tell, been systematized or manualized—there is no standard textbook account that is typically taught to our students, say—nonetheless we should be able to extract from what we might call philosophers' practical methodological wisdom that we can use to build what we can of an error profile of the method of cases. To do this, we will focus on places where there is at least some methodological guidance offered, and as such will leave out a lot of what has been written about the method of cases where these debates focus on the nature and status of "intuitions" or about how to best understand the underlying logical form of philosophical thought experiments (see, for discussion, Williamson 2007, Ichikawa 2009, and Malmgren 2011). As we've already said, while there is a lot of really good philosophy in these debates, they are typically not directed at methodological matters, and so while it is surely possible, at least in principle, that such debates could yield practical implications for the conduct of inquiry, we don't think that any have yet been found.

One can certainly find examples in our history of philosophers trying to work out error profiles for their own methods. For a lovely example, consider Austin's (1956) "plea for excuses." His methodological suggestions include: (i) preferring areas of language where ordinary practices are rich and well established, which is part of why he thinks excuses make a good target (one might think this a reason to give greater weight to, for example, verdicts about "knows" over verdicts about "justified"); (ii) avoiding areas of language upon which a lot of philosophical baggage has already accreted (perhaps a reason to prefer verdicts about "ignorant of p" over "knows that p"?); (iii) deploying substantively detailed settings for one's linguistic scenarios; (iv) dictionaries; and, perhaps more interestingly, (v) attending to usage of the term in specialist areas, which for his topic of excuses he would include law and psychology ("with which I include such studies as anthropology and animal behaviour" (p. 14)). What Austin is not including there, under the rubric of psychology, are methods like those of some experimental philosophers, doing empirical work on ordinary usage, but rather seeing what sorts of distinctions and nuances the psychologists have devised in response to their own explanatory needs on the topics of responsibility and action.

So let us look at what can be found in recent armchair practices along similar lines. The shared philosophical practical wisdom about the method of cases does indeed contain some methodological discussion of different sources of error when using case verdict evidence, as well as discussion about how these errors can be avoided or resolved. Here are some of the better-developed ones that we have been able to identify.

Narrative misunderstanding: One fairly tractable source of error involves misunderstanding the target case in some way. For example, untrained readers often misunderstand the disjunctive weakening steps in Gettier's original cases, so they fail to appreciate how the protagonists are meant to be justified in their beliefs. Another example, and one that has received a lot of attention recently, is the possibility that readers might fill in the details of cases in the wrong ways, since even fairly detailed scenarios will presuppose that readers round out the case with a huge range of unstated stipulations, very much along the same lines that readers engage with works of fiction (Ichikawa and Jarvis 2009).[1] These sorts of mistakes are usually weeded out by further reflection or deliberation, and because philosophers are generally sensitive to the difficulties of entertaining complex cases carefully, we frequently find philosophers glossing and unpacking their cases in some further detail in order to preempt such misunderstandings (Sosa 2007b).

Modal confusion: While the basic verdicts about a case are transparently accessible, philosophers generally acknowledge that more subtle aspects of

[1] Recall our discussion of how readers fill in the details of philosophical cases in Chapter 3.

those verdicts may not be. So, for example, philosophers may not have the same degree of access to the precise modal nature of these verdicts. Are they counterfactually true, merely possible, or necessary, and if necessary, nomologically, metaphysically, conceptually, or logically? And philosophers may not have any sense at all about what factors were responsible for the verdict going the way that it did. As Fodor (1998, 86–87) colorfully writes, "No doubt, intuitions deserve respect . . . [but] informants, oneself included, can be quite awful at saying what it is that drives their intuitions; sometimes it's just a fragment of underdone potato. This holds all the way from chicken sexing to judgements of grammaticality and modality. Good Quinean that I am, I think it is *always* up for grabs what an intuition is an intuition of." This is why case verdicts, while methodologically basic, will often occasion a great deal of downstream contemplation and argument, as philosophers try to figure out just what it is about a set of often closely similar cases that, nonetheless, elicited disparate verdicts.

Thus another fairly tractable source of error involves misidentifying the precise modal character of our case verdicts, which this modal opacity of case verdicts often makes a live possibility. Kripke (1980) famously discusses this confusion, warning that epistemic modality may be more permissive than metaphysical modality, and so we might mistakenly think that while water is actually H_2O, it nevertheless possibly could have been something else. Again, philosophers have suggested different ways of trying to weed out these kinds of mistakes. Bealer (1996), for example, recommends what he calls a *rephrasal strategy*. Bealer recommends that readers attempt to rephrase putatively errant case verdicts that it is possible that *A*, say, that water could have not been H_2O, into something like the form it is possible that a population of speakers in an epistemic situation qualitatively identical to ours would make a true statement by asserting 'A', but with appropriately different referents in 'A'. Bealer (2004) ends up rejecting the rephrasal strategy, for Burgean reasons, and he puts forward a different methodological strategy to discern where we may have mistaken a conflict between case verdicts that really are divergent in their modal force. He recommends that, instead of reporting case verdicts in terms of what is *possible*, we should report them in terms of what is *contingent*. The idea is that, although possibility admits of both epistemic and metaphysical readings, contingency is univocally metaphysical. So, if we will all report the case verdict that it is not contingent that water is H_2O, then there will be no difficulties here from a confusion between case verdicts of epistemic and metaphysical possibility. Strategies like these might give us reason for optimism, and we see this kind of optimism in Yablo (2006), who writes, "Carefully handled, conceivability evidence can be trusted, for if impossible *E* seems possible, then something else *F* is possible, such that we mistake the possibility of *F* for

that of *E*." The problem is that this works in practice only if we can discover and control from whatever distorting factors are responsible for causing us to mistake the possibility of *F* for that of *E*, and we simply do not have any general recipe for this kind of discovery and control. Yablo (1993) does not make it clear just what procedures we could adopt in order to be confident that we are handling our conceivability evidence sufficiently "carefully."

Pragmatic/semantic confusion: Another kind of mistake that we frequently make when thinking about philosophical cases involves confusing what would make a case verdict true or false and what would make that verdict felicitous or infelicitous. Again, philosophers have suggested ways of trying to weed out this kind of mistake. For example, Grice (1975) talks about something he calls *cancellability*: if what we think is funny about a proposed sentence is that it violates a *norm of conversational implicature*, then we should be able to deny that implication explicitly, and thereby render the sentence acceptable. For example, in a situation where there are ten philosophers in a room, each of whom is a metaphysician, is it correct to say, "Some philosophers in the room are metaphysicians"? Or does that statement actually entail that at least some of them are not metaphysicians? We can test this by considering a statement like "Some of the philosophers in the room are metaphysicians; indeed, all of them are." That statement seems kosher, and this is evidence that the initial negative reaction was not semantic in nature. (For an example of this kind of test that goes the other way, consider the unacceptability of "The number of philosophers in the room was odd; indeed, it was divisible by two.")

Having a "tin ear": Another kind of mistake that philosophers sometimes talk about when they talk about the mistakes made when using the method of cases is having a "tin ear," where this means, roughly, not being philosophically adept enough at discerning how philosophers most naturally describe a case or how best to map their own thoughts about the case. For example, Buckwalter and Stich (2014) attribute to Ned Block the observation that "not all that long ago it was a common practice for philosophers to dismiss people who didn't share their intuitions by saying that they have 'a tin ear'" (n36). And Williamson (2004) claims that "philosophers with a tin ear for natural language sometimes seem to misarticulate their own strong intuitions, using forms of words that do not express what they really want to say" (p. 121). And we can find various appeals here and there to the general possibility of philosophers having this kind of difficulty (e.g., Putnam 1990, 64). It is interesting to note that, while we find philosophers talking generally about this kind of error, it is hard to find actual token instances where philosophers level such accusations against specific rivals. What's more, there is very little guidance about how we might try to avoid *tinnitus philosophicus* or treat it when it has been diagnosed. It is also not clear

what should happen next, were two philosophers to have divergent case verdicts and each accuse the other of this malady.

Philosopher bias: A closely related kind of mistake involves the well-known *confirmation bias*, or what is called *experimenter bias* in the context of scientific methodology, and perhaps could thus be called *philosopher bias* here. Williamson (2007, 121) gives us a particularly nice presentation of this kind of worry:

> After all, philosophers defending a given position against opponents have a powerful vested interest in persuading themselves that the intuitions that directly or indirectly favor it are stronger than they actually are. The stronger those intuitions, the more those who appeal to them gain, both psychologically and professionally. Given what is known of human psychology, it would be astonishing if such vested interests did not manifest themselves in at least some degree of wishful thinking, some tendency to overestimate the strength of intuitions that help one's cause and underestimate the strength of those that hinder it. If one tries to compensate for such bias effects, one may be led to undercompensate or overcompensate; the standpoint of consciousness gives one no very privileged access to whether one has succeeded, for bias does not work by purely conscious processes. Its effects are much easier to observe in others than in oneself.

As Williamson acknowledges, being aware of this source of error is perhaps of little help in correcting for it, and there's been little by way of methodological guidance when it comes to philosophical bias, something made worse by the fact that merely being aware of this sort of bias can exacerbate its effects (Babcock and Loewenstein 1997).

In addition to these fairly well agreed upon features of the method of cases, there are other candidate views that are close to consensus views, but about which there are high-profile disagreements. So, for example, while many philosophers think that case verdicts should be preferred when they have been subjected to explicit and careful deliberation and reflection, some philosophers have recently argued that this kind of conscious cognitive effort would, in fact, potentially distort our actual case verdict, bending it toward our philosophical commitments (see, for discussion, Kornblith 2010, Talbot 2013, and Byrd 2021). And, to take another example, while many philosophers think that the subjective confidence we have in our case verdicts is a signal that those verdicts should be trusted in our philosophical reflections, some philosophers have recently argued that we should hesitate to give too much weight on subjective confidence (see, for discussion, Williamson 2004 and Saul 2007). One final example involves recent debates about whether verdicts about usual or esoteric cases should be significantly discounted in comparison with verdicts about more ordinary cases (see, for discussion, Fodor 1964, Kitcher 1978, Jackson 1998, Alexander

and Weinberg 2007, Williamson 2007, Cappelen 2012, Ichikawa and Jarvis 2013, and McKenna 2014). This all points to future potential for an error profile, but not one that has yet emerged. The armchair, and indeed the armchair plus the limited results of X-phi thus far, has not been able to settle these matters.

This isn't meant to be a comprehensive review of philosophical practical methodological wisdom about case verdicts; for example, we haven't said anything here about monitoring for scope ambiguities. But we think that it is comprehensive enough to give us a sense for the sorts of errors that philosophical folk wisdom reveals to us. For some of these errors, armchair philosophical practices seem to provide adequate resources to detect and correct them. For others, this is not so obvious. As a rough and unsurprising generalization, the error possibilities that are most successfully and directly addressed in philosophical practice are those for which a fundamentally philosophical accounting can be offered: errors that can be understood in terms familiar from logic, metaphysics, or the philosophy of language. But where the theoretical resources of armchair philosophy run out, so too, it would seem, does its methodological resources, which we hypothesize is why so little really *practical* wisdom is out there about things like tin ear and philosophical bias.

Even more to the point, the error profile that emerges from this philosophical folk wisdom simply fails to anticipate the kinds of problems that experimental philosophers have discovered.[2] It does not anticipate the extent to which different groups of people reach systematically different verdicts about the same philosophical cases or the extent to which the same people reach systematically different verdicts about the same philosophical cases when those cases are presented in different ways. That is, the error profile that emerges from this philosophical folk wisdom does not adequately address the extent to which case verdicts can fail to be robust. To the limited extent it is even somewhat aware of the existence of these problems, it offers no concrete guidance as to what to do about it. What's more, for reasons that we discussed in Chapter 1, this shouldn't be surprising. Subtle context effects will be undetectable to introspection or unsystematic observation, and demographic differences will often be obscured both by selection effects, where only those philosophers with the "right" verdicts will advance in specific domains of philosophical inquiry, and by the plain fact of professional philosophy's lack of demographic diversity. But while it perhaps shouldn't be surprising that philosophical folk wisdom lacks the methodological resources needed to detect these kinds of errors, it should be troubling, especially because experimental philosophy is still a very new area of research. And, so perhaps the most important lesson to be learned is not that experimental philosophy has revealed unexpected kinds of

[2] We have just been considering here the ways in which philosophers sometimes explicitly discuss sources of error, but one can also look at what is implicit in the way our practices are structured. For example, see our discussion of Joshua Knobe in Chapter 2, regarding the various ways in which our armchair practices seem not to have been in any way engineered to expect or manage these sorts of error possibilities.

errors, that is, kinds of errors that are not part of the error profile that philosophical folk wisdom gives us about the method of cases, but that it has revealed that philosophical folk wisdom just doesn't tell us very much at all about what really will or won't cause problems when philosophers use the method of cases.

Bringing this all together, not only does armchair psychology, by itself, provide a poor error profile for the method of cases, but the current practical methodological wisdom of the philosophical community also falls short. While it does provide something of an error profile, especially where the errors can be understood in philosophical terms, we can see that even in a few areas where it has become aware of threats of error, it has not been able to tell us what to do about them. And most of all, it is completely in the dark when it comes to managing the sorts of errors that experimental philosophy has brought to light.

2 Expertise to the Rescue?

Let's suppose that we are right that the error profile that emerges from our shared practical methodological wisdom about the method of cases does not anticipate the kinds of problems that have been discovered by experimental philosophers. There is at least one other place that one might hope to find already-in-place resources for error mitigation, namely, in our training and experience in issuing the case verdicts themselves. Perhaps we can manage these sorts of errors not based on what we know to watch out for, but simply in some sense, based on who we are, or at least who we have become by the time we are well along in our graduate education. Here are two particularly clear examples of this kind of defense of armchair philosophy:

> What is called for is the development of a discipline in which general expertise in the conduct of thought experiments is inculcated and in which expertise in different fields of conceptual inquiry is developed and refined. There is such a discipline. It is called philosophy. Philosophers are best suited by training and expertise to conduct thought experiments in their areas of expertise and to sort out the methodological and conceptual issues that arise in trying to get clear about the complex structure of concepts with which we confront the world. (Ludwig 2007, 151)

> Intuitions are and should be sensitive to education and training in the relevant domain. For example, the physical intuitions of professional scientists are much more trustworthy than those of undergraduates or random persons in a bus station. Scientists have and rely on physical intuitions, intuitions that are trained, educated, and informed and yet are good indicators of truth for those very reasons. In the same way, the modal intuitions of professional philosophers are much more reliable than either those of inexperienced students or the "folk." (Hales 2006, 171)

Let's call this the *expertise defense* of armchair philosophical practices.[3] The expertise defense is straightforward: while *some* case verdicts, namely, the verdicts arrived at by ordinary people thinking, perhaps for the first time, about philosophical cases, turn out to be problematically heterogeneous and unstable, we have reason to believe that *philosophers*' case verdicts will not. After all, philosophers have better concepts and theories, or at least a better understanding of the relevant concepts and theories, than ordinary people, have thought long(er) and hard(er) about these concepts and theories, and have been trained in how best to read and think about philosophical thought experiments that call upon us to apply these concepts and theories. Surely, all of this suggests that philosophers should be able to avoid these problems, which should make us less concerned about the fact that our armchair error profile for the method of cases did not, and could not, predict them.

Let's call this the *folk theory of philosophical expertise*. It is an attractive theory that promises to restore hope in at least some kinds of case verdicts. The trouble is that it turns out to be quite difficult to determine who has expertise about what and when. In general, only certain kinds of training help improve task performance and, even then, only for certain kinds of tasks, and there is reason to worry that philosophical training doesn't seem to be the right kind of training, nor does philosophical thought-experimenting seem to be the right kind of task (for discussion, see Weinberg et al. 2010). And so we will argue first that it is at best an open empirical question whether the folk theory of philosophical expertise is actually close enough to true to let appeals to philosophical expertise do the work of an error profile; and then turn to some specific empirical evidence suggesting that this is not so.

The Supposed Benefits of Philosophical Education

According to the folk theory of philosophical expertise, philosophical education involves thinking long and hard about philosophical issues, and philosophical expertise consists in having developed, through this process of critical reflection, increased conceptual competence and theoretical accuracy, as well as a special knack for reading and thinking about philosophical thought experiments that call upon us to exercise our conceptual competence and theoretical acumen. Here's a nice expression of this view:

> Theory-informed judgments in science may be more telling than the judgments of the uninformed because accurate background theory leads to more accurate

[3] Much of what follows draws on our discussion of philosophical expertise in Weinberg et al. 2010 and Alexander 2016. In addition to the papers discussed below, for additional discussion of the expertise defense, see Horvath and Wiegmann (2016), Löhr (2019), Seyedsayamdost (2019), Irikefe (2020), Wiegmann et al. (2020), Horvath and Wiegmann (2022), Schindler and Saint-Germier (2023).

> theory-informed judgment. The uninformed observer and the sophisticated scientist are each trying to capture an independently existing phenomenon, and accurate background theory aids in that task. Experts are better observers than the uninitiated. If the situation of philosophical theory construction is analogous, as I believe it is, then we should see philosophers as attempting to characterize, not their concepts, let alone the concepts of the folk, but certain extra-mental phenomena, such as knowledge, justification, the good, the right, and so on. The intuitions of professional philosophers are better in getting at these phenomena than the intuitions of the folk because philosophers have thought long and hard about the phenomena, and their concepts, if all is working as it should, come closer to accurately characterizing the phenomena under study than those of the naive. (Kornblith 2007, 35)

Let's begin with the idea that philosophers have better concepts. One way of cashing this idea out is to suggest that philosophical discussions typically involve technical concepts, and that philosophical education helps philosophers acquire these concepts. While there are certainly some philosophical discussions that involve technical concepts, for example, discussions about the nature of validity or warranted assertibility or grounding, the problem with this suggestion is that most philosophical discussions seem to involve rather more ordinary concepts, and for good reason. Since concerns about ordinary concepts are precisely what give rise to most philosophical discussions, if these discussions were then couched in purely technical terms, they would lose traction with the ordinary concerns that give rise to them in the first place (Knobe and Nichols 2008). It also seems to us that the case verdicts that get deployed in the literature most often concern ordinary concepts, such as *knowledge* or *moral permissibility*.

Let's assume this is right, and that most case verdicts that get deployed as evidence in our practices regard ordinary concepts. Maybe philosophers simply have a better understanding of these ordinary concepts, something that allows them to make more precise conceptual distinctions (Ludwig 2007). This would certainly make expert case judgments more evidentially valuable than folk case judgments, but it is important to be clear that questions of comparative conceptual competence are themselves open empirical questions and, even then, are extraordinarily hard to resolve (Knobe and Nichols 2008, Alexander, Mallon, and Weinberg 2010). Here's why. Evidence that philosophers have a *different* understanding of ordinary concepts would not be evidence that they have a *better* understanding of those concepts unless we have some independent reason to think that philosophical education somehow improves our conceptual understanding of ordinary, nontechnical concepts, and it is simply not clear how this is supposed to happen. Many philosophers seem to think that it happens through a process of trial-and-error learning, checking their conceptual judgments against some received standard. It is not clear, however, that we actually really do much of this sort of verdictive

fine-tuning in our professional training, as opposed to, say, simply disqualifying those who do not already share what might be spuriously set as "the" standard verdict for a case; given what we know now about variations in verdicts, there does not appear to be a good basis at present for disqualifying those who seem to have minority views on a case. Moreover, to the extent that we do, there is good reason to worry that we do not receive anything like the quantity of well-timed, objective feedback that the psychology of expertise tells us is necessary to improve conceptual understanding (for discussion, see Weinberg et al. 2010). Finally, for reasons related to our arguments against "reliability" in Chapter 3, we must observe that a concept that is "better" in terms of its baseline accuracy cannot simply thereby be expected to be "better" in the terms that matter here, of adequately immunizing its verdicts from variation, instability, or inconclusiveness.

While questions of comparative conceptual competence prove hard to answer, it seems easy to grant that philosophical education helps improve our understanding of philosophical theories. With this in mind, perhaps philosophical expertise consists in having mastered some set of philosophical theories or principles (Kornblith 2007, Ludwig 2007). The idea underwriting this suggestion is that philosophical theories can help shape our case verdicts, perhaps by making certain features of a given thought experiment salient or by guiding our interpretation of those features. This is an appealing suggestion, but one that faces a rather significant problem. Theoretical commitments are just as likely to contaminate our conceptual judgments as they are to decontaminate them. This means that the fact that expert case verdicts are theoretically informed does little to ensure that they are better than folk case verdicts. This point has prompted some philosophers to argue that we need to be careful that the theories that influence how we think about philosophical cases are accurate (Kornblith 2007), but the matter at hand is, precisely, that we have good reasons to think that armchair resources for philosophical theory-building are vulnerable to newfound sources of error. One cannot appeal to those very theories, then, in offering an account of what protects them from such errors. Finally, it is far from clear that there is enough even near-consensus agreement of philosophical theories to begin to provide such a shared standard to inform our verdictive capacities; we take up this issue of unresolvable philosophical dissensus at length in the next chapter.

Let's pause for a moment to take stock. On the view that we are examining, philosophical education involves thinking long and hard about philosophical issues, and this process of critical reflection is supposed to not only increase conceptual competence and theoretical accuracy, but also to improve our ability to read and think about the kinds of philosophical thought experiments that call upon us to exercise our conceptual competence and theoretical acumen. We have seen that questions of comparative conceptual competence and increased theoretical accuracy are hard to answer—both because they are tangled in unanswered philosophical questions themselves, and because they contain hard and yet-unanswered

empirical presuppositions—and even worse that there are reasons to worry that what answers we find will not bring good news to the folk theory of philosophical expertise. The situation actually gets even more complicated when we turn our attention to the idea that philosophical expertise involves some kind of *procedural knowledge*, or special "know-how" developed over the course of our philosophical education. Sosa (2009) provides a nice example of what this procedural knowledge might involve. The vignettes used in philosophical thought experiments require readers to import a certain amount of information not explicitly contained in the passages themselves, something that makes reading philosophical thought experiments similar to reading works of fiction.[4] If this is right, then we might think that, since philosophers have spent more time reading and thinking about philosophical thought experiments, they will be better able to get at the relevant details of a given vignette, better able to appropriately fill in details not explicitly contained in the vignette, and better able to entertain those details in their imaginations. Williamson (2007) advances a similar position, arguing that philosophical thought experiments involve deductively valid arguments with counterfactual premises, and that evaluating them requires a mixture of imaginative simulation, background information, and logic. If this is right, then we might again think that, since philosophers have spent more time reading and thinking about philosophical thought experiments, they will be better able to pick out the relevant details of a particular vignette, better able to engage in counterfactual reasoning, and better able to make appropriate logical inferences.

The idea that philosophical expertise consists of some kind of special procedural knowledge is certainly attractive, but it is important to be clear that questions of comparative procedural expertise, like questions of comparative conceptual competence and increased theoretical acumen, are open empirical questions. As we argued in the previous section, any such procedural knowledge here is not spelled out clearly in the practical methodological wisdom of the field. To the extent that this question turns on what is going on in philosophers' heads without their articulating it, it simply cannot be answered without careful empirical investigation.

One last idea of expertise that seems a part of the folk philosophical theory is that our simply being better at cognitive reflection, and generally being much more willing to engage in it, should be expected to give us improved performance at producing verdicts.[5] This idea about the relationship between reflection and cognition is what makes it seem so natural to think that, since philosophers spend more time thinking about philosophical issues, expert philosophical cognition should be better than folk philosophical cognition. The problem is that the relationship

[4] See our discussion in section three of Chapter 3.

[5] In the context of the metaphilosophical debate about the method of cases, this idea has come to be known as the *reflection defense* (Weinberg et al. 2012). For additional discussion see Kauppinen (2007), Ludwig (2007), Liao (2008), Horvath (2010), Nado (2015), Hannon (2018), Kneer et al. (2022), and Byrd (2023a).

between reflection and cognition is not this straightforward. There are times when reflection helps improve philosophical cognition. Goldman (2007) provides some nice examples: reflection can help us realize that we have been misinformed or uninformed about some relevant details of a particular case, that we had lost track of some of the relevant details, or that our initial judgments about what details are relevant were contaminated by our theoretical commitments. But there are also times where reflection serves as an echo chamber, simply ratifying whatever initial judgments we might have made, and increasing the confidence that we have in those judgments without increasing their reliability (Kornblith 2010, Weinberg and Alexander 2014).[6] Similarly, while reflection may sometimes eliminate a context effect, it also has the power to lock such an effect in and exacerbate it through the power of motivated cognition. All in all, again we find that extensive scientific work would need to be done to establish any claim about the comparative merits of expert philosophical cognition and folk philosophical cognition, and in particular that the former is sufficiently shielded from the sorts of errors under consideration here.

Empirical Work on Expert Case Verdicts

So far, we have focused on the perceived benefits of philosophical education, and found that the picture is not quite as clear as we might have hoped, something that puts pressure on our folk theory of philosophical expertise. We have also noticed that many of the questions that have been raised are open empirical questions: questions about comparative conceptual competence and improved theoretical acumen, as well as questions about the role that philosophical education might play in the production of special procedural knowledge. But the empirical questions are not totally open, as in fact a number of studies on philosophical expertise have already been conducted. And on the whole, they do not tell a story that is friendly to the armchair.

If philosophical education produces genuine philosophical expertise, then we might expect expert case verdicts not to display the same patterns of problematic sensitivity that folk case verdicts display, and that philosophical education does not introduce new patterns of problematic sensitivity. Let's see whether this is the case. As we discussed in Chapter 1, some folk case verdicts are *unstable*, including folk case verdicts about what actions are morally good/bad. In particular, it seems that folk case verdicts about what actions are morally good/bad in one case and

[6] Nick Byrd has attempted, over a series of papers, to carefully set out what it would take to actually test whether or not, and if so, in what ways, reflection might improve philosophical case verdicts. See, e.g., Byrd (2021, 2022, 2023a, 2023b), Byrd et al. (2023).

what actions are morally good/bad in another case depend on the order in which the cases are presented. With that in mind, consider the following two cases:

> Jane is standing on a footbridge over the railroad tracks when she notices an empty boxcar rolling out of control. It is moving so fast that anyone it hits will die. Ahead on the track are five people. There is a person standing near Jane on the footbridge, and he weighs enough that the boxcar would slow down if it hit him. (Jane does not weigh enough to slow down the boxcar.) The footbridge spans that main track. If Jane does nothing, the boxcar will hit the five people on the track. If Jane pushes the one person, that one person will fall on the track, where the boxcar will hit the one person, stop because of the one person, and not hit the five people on the track.
>
> Vicki is standing by the railroad tracks when she notices an empty boxcar rolling out of control. It is moving so fast that anyone it hits will die. Ahead on the main track are five people. There is one person standing on a side track that doesn't the main track. If Vicki does nothing, the boxcar will hit the five people on the main track. If Vicki flips a switch next to her, it will divert the boxcar to the side track where it will hit the one person, and not hit the five people on the main track.

If expert case verdicts about what actions are morally good/bad are influenced by *stable* moral considerations, then these verdicts should not be affected by the order of presentation. But this is not what experimental philosophers have found. In a study involving four different groups (ethicists, philosophers, academic non-philosophers, and non-academics), Eric Schwitzgebel and Fiery Cushman (2012) found that everyone's verdicts about the moral valence of the relevant actions in these two cases were affected by the order in which they were presented. They grouped responses into two categories (*equivalent responses*, where evaluations of moral valence were identical across the two cases; and *inequivalent responses*, where participants judged the relevant action in the first to be morally worse than the corresponding action in the second case) and found that participants, regardless of academic background or experience, were more likely to give equivalent responses when the first case was presented before the second case than they were when the second case was presented before the first case. This suggests that at least some expert case verdicts about whether an action is morally good/bad are influenced by the order of presentation.

Let's look at another example. Folk verdicts about a variety of philosophical issues seem to be influenced by normative considerations (for discussion, see Knobe 2003, 2010 and Cova 2016). The *side-effect effect* is probably the most famous example of the influence that normative considerations have on how we ordinarily think about the world. Knobe (2003) found that ordinary people are considerably more inclined to judge that an agent brought about a side-effect intentionally when they regard that side-effect as morally bad than when they regard it as morally good.

Knobe's model has also been used to show that normative considerations influence a wide variety of other folk psychological judgments, including judgments about advocacy, causation, choice, decision, desire, knowledge, and preference. We are just beginning to see how widespread this influence might be and to come to terms with what this is telling us about how our minds work and how we ordinarily understand the world around us. With this in mind, consider the following two cases:

> A baby was born with a rare genetic condition. The doctors told the baby's parents: "If this baby drinks its mother's milk during the first two weeks of life, it will grow up to have extraordinary mental abilities that make it able to solve very complicated math problems. However, if you instead give it this expensive formula we sometimes use, it won't develop the extraordinary abilities and will just be normal." The parents said: "We have decided not to give the baby the expensive formula. We will just be feeding it with its mother's milk." As expected, the baby grew up to have extraordinary mental abilities that made it able to solve very complicated math problems.
>
> A baby was born with a rare genetic condition. The doctors told the baby's parents: "If this baby drinks its mother's milk during the first two weeks of life, it will grow up to have serious psychological disabilities that will make it unable to solve even very simple math problems. However, if you instead give it this expensive formula we sometimes use, it won't develop the extraordinary abilities and will just be normal." The parents said: "We have decided not to give the baby the expensive formula. We will just be feeding it with its mother's milk." As expected, the baby grew up to have serious psychological disabilities that will make it unable to solve even very simple math problems.

If expert case verdicts about whether a specific trait is innate are influenced by purely scientific considerations, then these verdicts should not be sensitive to normative considerations. This is not what we find, however. In a study involving both academic professionals and people working outside of academia, or, more particularly, outside of philosophy and specific scientific disciplines, Knobe and Samuels (2013) found that people are considerably more inclined to judge that a trait is innate when the expression of that trait depends on actions they regard as morally good than on actions that they regard as morally bad, and that this is not affected by profession or level of professional training. This suggests that expert philosophical case verdicts and folk philosophical case verdicts share sensitivity to normative considerations, although it remains an open empirical question whether these kinds of considerations form part of our conceptual competence or simply figure into our conceptual performance (for discussion, see Alexander, Mallon, and Weinberg 2010).

Taken together, these results suggest that at least some expert philosophical case verdicts display the same patterns of problematic sensitivity that are displayed by

folk philosophical case verdicts, something that is bad news for our folk theory of philosophical expertise.[7] The news actually gets worse when we consider whether education might actually introduce new patterns of problematic sensitivity. Consider the following case:

> Ivy is a high school student in Hong Kong. In her astronomy class, she was taught that Tsu Ch'ung Chih was the man who first determined the precise time of the summer and winter solstices. But, like all her classmates, this is the only thing she has heard about Tsu Ch'ung Chih. Now suppose that Tsu Ch'ung Chih did not really make this discovery. He stole it from an astronomer who died soon after making the discovery. But the theft remained entirely undetected and Tsu Ch'ung Chih became famous for the discovery of the precise times of the solstices. Everybody is like Ivy in this respect; the claim that Tsu Ch'ung Chih determined the solstice times is the only thing that people have ever heard about him.

In a study involving professional philosophers and linguists, Machery (2012) found that expert case verdicts about reference, in particular, expert verdicts about Ivy's referent when she uses the name "Tsu Ch'ung Chih," were influenced by a person's area of research specialization. People with certain areas of research specialization (e.g., semantics and the philosophy of language) were more likely to have Kripkean verdicts than people with other areas of research specialization (e.g., discourse analysis, historical linguistics, and sociolinguistics). This suggests that educational background influences at least some philosophical case verdicts about reference, something that would be welcome were the influence consistent; but, as Machery argues, these studies suggest an inconsistent influence of educational background on expert philosophical case verdicts, something that should give us further reason to worry about the supposed benefits of philosophical education.

We began with the attractive idea that philosophical education involves thinking long and hard about philosophical issues, and that philosophical expertise consists in having developed, through this process of critical reflection, increased conceptual competence and theoretical acumen, as well as a special knack for reading and thinking about philosophical thought experiments that call upon us to exercise our conceptual competence and theoretical acumen. And we noted that it is an open question whether this folk theory of philosophical expertise can restore hope in the value of verdictive evidence. We saw that questions about comparative conceptual competence and improved theoretical acumen, as well as questions about the role that philosophical education might play in the production of special procedural knowledge, are actually empirical questions,

[7] See, for additional discussion, Horvath and Wiegmann (2022), who studied whether ethicists are immune from several different kinds of the classic framing effects studied by Tversky and Kahneman (1981).

and suggested that the best place to look for answers to these questions is science, in particular, recent empirical work on the nature of expert case verdicts. And we saw that this work suggests that at least some expert case verdicts display the same patterns of problematic sensitivity that are displayed by folk case verdicts, and that educational background might introduce new kinds of problematic verdictive sensitivity. While this is bad news for the folk theory of philosophical expertise, it is important to note that these are still early days. We are just now beginning to understand philosophical cognition, and some recent empirical work highlights ways for improving folk philosophical cognition, something that suggests how we might achieve genuinely expert philosophical cognition (see, e.g., Turri 2013). This is an exciting development that highlights the fact that the more we learn about philosophical cognition, the more there is to learn, and that underscores the importance of continued experimental work on philosophical cognition and philosophical expertise.

3 The Methodological Irrationality of Armchair Philosophy

With all of this in mind, we are ready to present our big-picture take on what's really wrong with armchair philosophy. First, the kind of experimental work rehearsed in Chapter 1, and discussed extensively and systematically in Machery (2017), demonstrates that the method of cases is vulnerable to a wide range of errors that could not be predicted from the armchair; indeed, so wide a range of errors that we have every reason to believe that there are many more empirically live threats of error that have yet to be discovered. As we have just now discussed, these kinds of errors are not part of the error profile for the method of cases as can be found in our profession's practical methodological wisdom; nor do we have much reason to expect that our philosophical training and education gives us much help in managing these threats of error. Pitched in terms of methodological rationality, all of this means that the community of analytic philosophers will likely be *methodologically irrational* should we continue to use the method of cases from our armchairs. The error profile we get from our armchairs does not predict the kinds of errors we are likely to make, and the resources available from the armchair do not help us mitigate those errors. There are substantial benefits to inquiry to be gained from achieving a better error profile, and in order for us to secure those benefits, something will have to change.

This is a very different kind of worry than the kinds of worries that philosophers have raised using some version of the impeachment argument schema that we discussed in Chapter 4. The problem with how analytic philosophers ordinarily use the method of cases is not that philosophical case verdicts have some epistemic property that makes them incapable of providing evidence. The problem is that the way that analytic philosophers ordinarily use the method of cases leaves them without the resources needed to predict the kinds of problems that present a live

possibility when using the method of cases, and unable to supplement that method in a way that could help mitigate their substantial risks of error. What's really wrong with how analytic philosophers use case verdicts as philosophical evidence is not anything especially or distinctively wrong with the case verdicts, but rather with the manner of use, that is, stuck in the armchair.

We are looking to argue that the only methodologically rational response for the community to adopt is a broad investment in experimental philosophy. Since methodological rationality is fundamentally about gaining benefits that outweigh costs, we still need to say more about the benefits and costs that would be associated with redirecting some of our community resources toward better error prediction and error mitigation, and, in particular, toward experimental philosophy. These benefits will be measured in terms of the long-term acquisition of novel truths, and these costs can be measured in terms of resources like the time and effort that philosophers give to different kinds of first-order or metaphilosophical projects, and perhaps other things like academic positions, space in academic journals, and so on. The final two chapters will aim to play up such benefits of a greater investment in X-phi, and to offer reassurance as to the actually fairly modest costs. Before we can do that, though, we need to address head-on a different kind of objection to our deployment of MR. For there are a range of unusual circumstances in which, even when a research community seems to be badly underinvested in error mitigation, it is nonetheless still methodologically rational for them *not* to engage in a diversion of resources. We will contend that none of these circumstances apply to the method of cases today.

So, what are some of the situations in which it would still *not* be rational to move some community resources toward error mitigation, even when new errors and error possibilities have been discovered? One reason, which we can dispatch rather quickly given what we've already said, would be that it turns out the error possibilities may be newly discovered but not newly problematic; that is, should they already turn out to be anticipated in a community's practices, by some cosmic stroke of methodological luck. But as we argued above, the sorts of errors that experimental philosophy reveals are not ones that have been, or even could be, anticipated by the philosophical folk wisdom that has grown up around armchair philosophical practices or philosophical education and training.

Another reason would be if the community simply cannot identify any good avenues for investing in error mitigation. For example, an obvious source of error in classics is that we only have a fairly limited sample of ancient texts available to us, so we are at a constant risk of overgeneralizing from whatever works and fragments that have happened to have survived. But there is just no option to invest resources in, say, a time machine in order to go back and rescue more texts. (Though great resources will be spent in rescuing the very occasional text that can, such as with recent work using advanced computational techniques to "read" some Vesuvius-charred scrolls found at Herculaneum.) An MR-based argument for classicists to

devote more resources to mitigating this one kind of error will thus literally go nowhere. But armchair philosophy is in no such situation. Obviously the profession can afford to do *some* X-phi, because it already does so; and quite obviously analytic philosophy *could* invest in at least some *more*; and, most of all, for many of the sorts of errors that experimental philosophy reveals, *it can often also help manage them*. At a minimum it can make us aware of many of the vulnerabilities in the case verdicts, and allow us to steer clear from particularly fraught kinds of materials. For example, it seems like we have enough evidence already to suggest that epistemologists should cease using fake barn cases as examples of non-knowledge (Colaço et al. 2014). But it also offers resources to compensate or correct for such sources of error. For example, experimental techniques allow us to control for order effects using techniques that cannot be pursued from the armchair, such as presenting the cases in different orders to different participants and then averaging across the different orders; or to look actively for the presence of demographic variability, by taking the required steps to recruit a broad sample. And quantitative modes of inference allow for further ways of making less error-fragile inferences than those which are still largely in use in analytic philosophy. And importantly, the adoption of these techniques does not require any impossibly extravagant expenditure of resources. We don't need CERN to build us a large-scale armchair accelerator.

A third reason that a community might rationally refuse to engage in a trade-off of resources of the sort we are suggesting is when the *opportunity costs* associated with the investment of community resources in error mitigation could still be too great. That is to say, even when we have good reasons to expect that mitigating errors would be a high-value investment, and good reasons for thinking that an investment in error mitigation is one that we could make, it might still be methodologically irrational to make such investments when those investments would have to come at the cost of not being able to make other, comparably high-value investments. (Or at least, under such circumstances it would *not* be methodologically irrational *not* to make such investments.) We must consider what the opportunity costs of any such transfer of resources might be, and whether the sacrifice would be so costly as to render dubious the application of MR to motivate that transfer.

Here are three ways in which the opportunity costs could involve this sort of problematically high-value sacrifice. We will argue that none of them seem to apply to the case at hand.[8]

The first, and perhaps most obvious, way that opportunity costs could involve high-value sacrifice is when the costs come from *materially non-substitutable* research activities. Imagine, for example, that we could mitigate certain kinds of computation errors associated with using mechanical calculators to assist in the

[8] These are all the ones we have been able to devise, but we have to note that, if a clever interlocutor can devise a further kind of situation, and cogently argue that it applies to the circumstances of armchair philosophy today, we would need to attend carefully to it.

performance of sufficiently complicated computations only at the expense of giving up the use of mechanical calculators altogether. It is rather easy to see that mitigation efforts that required the complete abandonment of these devices would be a bad proposal, and not the least of all because there are large-scale numerical questions whose answers would be put practically out of reach altogether if we had to go back to relying on "calculators" in the old-school sense of professional human number-crunchers. Even if *ex hypothesi* our mechanical calculators don't catch these new sorts of errors, they do provide evidence of many arithmetical truths that we could not get without some such device. It is unclear how the community should proceed in such a case, but ideally they would find some other strategies for error mitigation that would not require this total surrender of the technology.

For a more realistic example, consider the role of particle accelerators like the Large Hadron Collider in fundamental physics. A key MR reason for the enormous investment of resources in such devices is that they provide evidence of a sort that no other device can. The same is true for extraterrestrial telescopy like the James Webb Space Telescope, or in the social sciences, for the kind of nation-level gathering of data represented by the Census. An MR-based proposal for resource reallocation that would substantially deprive us of such non-substitutable sorts of evidence is going to be a hard proposal to defend.[9] But plainly, philosophy has no such materially non-substitutable research activities. There are no highly concentrated expenditures of resources that gain us access to truths that could be had in no other way. We are, by and large, an immaterial discipline, for better or for worse. Even armchairs themselves, in their literal manifestations as furnishings and not just used as a metaphilosophical metaphor, play at best a highly indirect and, as it were, supporting role in philosophical knowledge production. This is no mark at all against philosophy, of course. It just means that we don't need to worry about this sort of opportunity cost problem when considering whether we are in a state of methodological irrationality.

A second, diachronic way in which opportunity costs might be problematically high is when they'd force us to give up some other *temporally irrecoverable* research investments. It's crucial to keep in mind that methodological rationality is about how best to expend our resources to get a hold of novel truths in the long run, and as such we will not worry so much about short-term opportunity costs if they can be made up for over time. Temporally irrecoverable opportunity costs are those that cannot be compensated for later, where no expenditure of resources at a later time can yield the evidentiary results that such an expenditure could achieve now. For example, if an asteroid were to pass close enough to the earth to allow a

[9] Such cases also are significant for not just their high overall evidential value, but their high value at the margins. Moving from a particle accelerator that is 5% complete to one that is 98% complete has nowhere near the value associated with getting that last 2% of the construction project finished. We wouldn't want to sacrifice the collider, and to sacrifice even a nontrivial part of the apparatus may be the same as sacrificing the whole.

manned exploration mission, never to pass this way again, then foregoing such a mission might be enormously costly, in MR terms. Here, again, this possibility just does not seem to be one that philosophers are likely to have to face. Consider the opportunity costs of some substantial number of philosopher hours of traditional research being given up to make room for more investment in error mitigation. Let's stipulate that this means that at least some philosophical insights that would have been achieved by time *t* would, as a consequence of the reallocation of resources, fail to be achieved at that time. Now, as we've already said, if we have reason to worry that these insights will be lost and gone forever, then that would be a problematic sort of opportunity cost. And under such conditions we would be in danger of it becoming unclear how to compare that hypothetical potential cost to the hypothetical potential gains resulting from the reallocation of resources. But philosophers just don't face this kind of problem. We simply should not expect that any such foregone insights would be an eternal loss; at worst they would be postponed—instead of *t*, they will occur at *t'*, and even if *t'* is years later than *t*, in the long run that won't be a meaningful difference. The plain fact is that philosophical research activity is rarely, if ever, temporally irrecoverable.[10] Resources that are transferred away from first-order research to error mitigation will at worst mean insights delayed, not insights abandoned. And one thing to very much keep in mind: catching errors earlier rather than later carries its own efficiencies. Suppose we ultimately learn that, say, there are no philosophically meaningful stakes effects on knowledge verdicts.[11] Think of how much time and energy we collectively would have saved had we learned so twenty years earlier!

Conditions of *extreme methodological poverty* present a third, if rather more hypothetical, way in which the opportunity costs could be problematically high value. Consider this example, adapted from Jessica Brown's paper, "Intuitions, evidence, and hopefulness" (Brown 2013):[12]

> Suppose that a certain company is the only one in the world that produces a reliable measuring instrument of a certain kind, and they have a very well-developed account of the error profile of that instrument—let's call it the *archeometer*. This company has been part of our research community, meaning that members of

[10] One may object that some topics and projects may be temporally urgent because of their salience to current matters such as current debates about human rights, or the threat of incipient fascism. This strikes us as highly plausible, and to the extent that it is so, it seems clear to us that our arguments would militate against transferring resources out of those specific research areas. Nonetheless, it is patently the case that such researches make up a fairly small fraction of current philosophical research.

[11] We will discuss this more in the next chapter.

[12] To be clear, Brown presents her original version as a challenge to the argument from hopelessness that we discussed above. We have reframed her case here so that it is a challenge instead to our claim here about when resources should be reallocated to avoid underinvestment in error mitigation, but please note that Brown's original thought experiment seems to us to strike home against its intended target, and she is in no way responsible for our adaptation of her materials, for argumentative purposes not her own.

> the company, and in particular, those members who have this well-developed account of what sorts of errors the archeometer is prone to and how to handle them, are members of our research community. So our collective practice with the archeometer will not likely be in a state of underinvestment in error mitigation, even if the people with this special knowledge of the archeometer keep that knowledge to themselves, perhaps because it is a trade secret. But suppose that there is a catastrophe which tilts the balance; one that wipes out the company and all of its specialists, and so our research community loses their knowledge of how the archeometer functions, and how to manage its errors. It would take a great deal of time, money, and personnel to recover anything like the knowledge that the company's specialists had possessed about the archeometer, the recovery of which would be necessary for robust error mitigation. What should the post-catastrophe research community do when it comes across an archeometer stamped with the name of this impeccable firm? The current community has no means of detecting and correcting errors when they use their archeometers, but nonetheless they have exceedingly good information that their use of them is highly (if still only moderately) reliable. One last stipulation: as a further consequence of the catastrophe, our resources for conducting research are now scarce and there are many other instruments that have been lost altogether, but which could be reinvented with a nontrivial but practicable expenditure of resources. Finally, right now there is no other instrument that can perform the same measurements as an archeometer. It thus follows that any resources spent trying to rediscover the error profile of the archeometer could be spent with much greater methodological yield on reinventing these other, now-lost devices. Surely, shouldn't this community continue to use their archeometers, even without substantially reallocating any resources to archeometric error mitigation at this time?

We believe that such conditions *would* be ones in which investment in error mitigation might well be outweighed by the value of investing in other, now-absent means of first-order investigation. We thus admit that it would be methodologically rational for the researcher community in this hypothetical case to refrain from the kinds of resource reallocation that we have in mind under the kinds of extraordinarily special circumstances described in the thought experiment. But this admission makes it all the more important to point out that what makes special circumstances special is that they are, well, truly *special*. It's not just a stylistic flourish in Brown's case, or our adaptation of her case, that we are to imagine a "catastrophe" of some sort that has reduced us to methodologically impoverished circumstances. After all, by and large, well-developed research communities tend to have lots of resources. Of course, there are lots of demands on those resources too, but not so much so that *any* such reallocation would be overly painful.

With all of this in mind, we think that we can integrate Brown's insight about this kind of case into our MR framework simply by pointing out that what she has

done is identify a kind of situation—namely, extreme methodological poverty—where it may well not be methodologically rational to try to correct for apparent error-mitigation underinvestments. If the resource budget for our inquiries on the whole is parlously low, then it is likely that most methodological goods that we are investing in are still at such an early point on their veritistic utility function, and one that has a sharply positive first derivative, so even small sacrifices may be keenly felt. (Compare how much value $20 more—or less—has when you are just scraping by on a grad student stipend.) Moreover, for all that we have emphasized the value of error mitigation for our methods, the amount of actual output from any such method imposes a certain kind of ceiling on the value of error mitigation for it. Being able to catch all the errors in some instrument will be of dubious value if no one can ever afford to run the instrument in the first place.

Here again we see a kind of in-principle exception that clearly does not apply to analytic philosophy. For all that we would obviously like to receive a greater share of society's resources to spend on our beloved discipline, nonetheless, philosophy is just not remotely in such a state of profound poverty of the resources of inquiry. There are thousands of working philosophers each devoting hundreds of hours every year to philosophical research. We also do not seem to find ourselves in general at a point currently where the marginal value of our collective research is especially high, such that the value of gaining or losing a bit of resources for our investigations would accordingly yield or sacrifice a great deal of acquirable truth. Just to be very clear, this state of affairs does not at all reflect poorly on philosophy! What we are saying here of philosophy we take to likely be true across the great majority of academic disciplines. We take it that the sort of radical methodological impoverishment described in Brown's thought experiment is an extreme one, and an extremely rare one in modern times.

Having said that, it also seems possible that even amid a *general* state of at least minimally adequate resources, there could be pockets of *local* impoverishment. Suppose that there were a research sub-community within philosophy for which there was such a low amount of support that it was practically as if a Brown-style scenario had befallen those researchers. It could then be the case that transferring resources away from the research on those topics would not in fact be methodologically rational; or at least, it might not be at all clear what methodological rationality would require of such a case. We can think of two ways in which this could, with any plausibility, occur within philosophy. First, there could be research areas so nascent that they have had no time to garner much support or build an adequate sustaining professional infrastructure. Second, there could be research areas that are not currently so profoundly under-resourced, but which would become so if forced to absorb the bulk of the redistribution. That is, if we were going to have to propose basically shutting down some existing areas of philosophical inquiry in order to free up resources for error mitigation, then that could present a localized version of the third possibility just explored. Unlike with the first three

possible exceptional circumstances described already, which patently do not apply to the circumstances of current philosophical practice, the legitimate possibility of local impoverishment places nontrivial constraints on how any such redistribution of resources is carried out.

Fortunately, they are not hard constraints to satisfy. For the costs will generally be borne in a highly distributed way, coming in bits and pieces from all across the philosophical community. It's not like we will be funding X-phi error-mitigation projects by putting all the other subfields into a hat, pulling a name out and saying, "Ok, scholarship on [opens envelope] Humean virtue ethics gets the axe in order to make way for experimental philosophy!"[13] Moreover, where there are more recently rising subfields where we can recognize that even small amounts of resources can be expected to produce high yields, we simply ought to exclude them from possible reduction.[14] Indeed, for at least some such subfields, an increased investment in experimental philosophy might well involve an increased investment in that very area, for X-phi work to be done within that area—keep in mind, experimental philosophy is not a distinct domain within philosophy so much as a distinct way of doing philosophy across domains.[15]

With all of this in mind, it seems to us that philosophy simply is not an exception to the general rule that, when research communities face unexpected problems that their methods cannot mitigate, those communities are rationally obligated to invest substantially in developing resources to predict and manage them. And this paves the way for the kind of methodological reform that we are calling for here.

[13] One possible exception: it may be that any such redistribution will be somewhat disproportionately felt by mathematical logic, since part of what we will advocate in Chapter 7 is altering the formal methods requirements in analytic graduate programs so that not everyone will be taught, say, how to prove the first and second incompleteness theorems or the upward and downward Löwenheim–Skolem Theorem. Instead some students will receive some formal training in empirical methods. But it seems to us that already many such graduate logic courses are covered by faculty who are not AOS mathematical or philosophical logic.

[14] There are legitimate concerns to be raised here about whether the burden of such redistributions may fall disproportionately on marginalized areas within philosophy, or for that matter on philosophers who are themselves members of marginalized groups. Such political and ethical questions seem to us a mostly distinct question from the one at hand here, within the MR framework.

[15] The kinds of problems that we have been looking at up to this point focus on conditions internal to a research community's resource balance sheet. We want to distinguish these sorts of pressures from another kind of pressure that can be placed on the arguments that we have been making about resource reallocation. This kind of pressure has to do with demands on a research community's resources that are external to the concerns for inquiry itself. We do not think that these kinds of external pressures present a *counterexample* to our MR approach, but we do think that they are cases that involve practical or ethical considerations that *override* considerations of methodological rationality. In these kinds of cases, a research community's violation of methodological rationality would be exculpable because there are stronger normative considerations in play that override the normative considerations associated with methodological rationality. Suppose, for example, that researchers are facing a global pandemic and need a vaccine developed *now*, and so they cut a lot of research corners that they wouldn't normally cut in the course of careful biomedical research. We are inclined to say that these researchers may possibly have made very much the right decision, all things considered, to cut those corners, depending on what those corners happen to be and how it all works out, of course. But to use that lingo seems to us to imply that the relevant norms, that is, the "corners" in question, were not rendered inert, but were overridden.

4 Methodological Rationality and Experimental Philosophy

It is striking that when experimental philosophers have gone looking for trouble, they have not had too much difficulty finding it, and that our armchair practices just don't seem to be up to the task of consistently predicting where the trouble lies or providing the resources needed to resolve it. At this point, it is natural to ask, are we sure that, in pursuing these sorts of methodological worries in this way, we aren't simply looking to abandon philosophy as we know it? Interestingly, we can find the beginning of a response to this worry from perhaps an unlikely source. Timothy Williamson (2007), who spends much of *The Philosophy of Philosophy* defending armchair philosophical practices, nevertheless ends the book by suggesting that philosophers "must do better." He has in mind ways in which current philosophical practices can go astray because of linguistic carelessness, and argues that philosophers should attend closely to linguistic aspects of our philosophical activities, on the model of how scientists must understand the tools that they deploy in their investigations:

> Philosophers who refuse to bother about semantics, on the grounds that they want to study the non-linguistic world, not our talk about the world, resemble scientists who refuse to bother about the theory of their instruments, on the grounds that they want to study the world, not our observation of it. Such an attitude may be good enough for amateurs; applied to more advanced inquiries, it produces crude errors. Those metaphysicians who ignore language in order not to project it onto the world are the very ones most likely to fall into just that fallacy, because their carelessness with the structure of the language in which they reason makes them insensitive to subtle differences between valid and invalid reasoning. (pp. 284–285)

The point is straightforward, and we hereby extend it *mutatis mutandis* to these recent metaphilosophical debates about the armchair method of cases:

> Philosophers who refuse to bother about what shapes our thinking about philosophical cases, on the grounds that they want to study philosophical questions, and not the methods that we use to try to answer them, resemble philosophers who refuse to bother about semantics, on the grounds that they want to study philosophical questions, and not how people talk about them. Such an attitude might be good enough for amateurs; applied to more advanced inquiries, it produces crude errors. Those philosophers who ignore empirical work that sheds light on the method of cases in order to preserve the ideal of methodological self-sufficiency are the very ones mostly likely to fall into error, because their carelessness of the shape and structure of the method of cases makes them insensitive to

> subtle differences between problematic and unproblematic verdicts about philosophical cases.[16]

As Williamson also points out, and as we discussed at length in Chapter 3 where we talked about the error fragility of philosophical practice, "errors easily multiply to send inquiry into completely the wrong direction" (p. 288). This underscores the fact that the kind of unsystematic guesswork that makes up both philosophical folk wisdom about the method of cases and folk theories of philosophical expertise should be no more acceptable to philosophers than unsystematic guesswork about language and logic would be. And just as philosophers exercise appropriate methodological care when they attend carefully to logic and language, methodological rationality calls for philosophers to attend carefully to empirical work that sheds light on the method of cases—that is, experimental philosophy!

In the next two chapters, we argue that the benefits of adding experimental tools and methods to the set of resources that are available to analytic philosophers not only does not involve the kinds of radical changes to the shape and character of professional philosophy that analytic philosophers have often feared it does, it also provides resources that analytic philosophers can use to make substantial progress on the kinds of questions that we have been trying to answer all along. But we want to close this chapter by talking about another way that experimental tools and methods can aid in error mitigation, namely, by helping remove the *inquirers themselves* as sources of error—especially when effects are present but inconclusive. Human cognition just cannot help resolving ambiguous evidence in favor of a preferred hypothesis, something that psychologists call *myside bias* (Mercier 2017, Stanovich 2021). Myside bias is the widespread and hard-to-remove tendency for people to "evaluate evidence, generate evidence, and test hypotheses in a manner biased toward their own prior beliefs, opinions, and attitudes" (Stanovich et al. 2013, 259). Here we will focus on people's tendency to evaluate ambiguous signals in favor of their preferred hypotheses, as discussed a bit above as "philosopher bias." One way in which philosophers might do this when using the method of cases would be to mentally "round up" or "round down" the experienced strength of their own inclinations toward particular case verdicts, according to how well those verdicts fit, or problematize, their own preferred philosophical theories. As we will discuss in the next chapter, philosophers sometimes find it all too easy to give some sort of preference to presentations of a thought experiment that yield the "right" results for them, and *mutatis mutandis* for their philosophical interlocutors. And it is not at all surprising that philosophers do this. We *are* people after all, even if we are somewhat weird ones (or WEIRD ones). What's more, while philosophers are generally *smart* people, one of the most interesting empirical

[16] This section is borrowed from Weinberg (2009, 463).

results about myside bias is that people who score high on intelligence tests and other measures of cognitive ability are just as likely as anyone else to engage in this kind of biased thinking. Being smart and thinking hard have proved to be of very limited help here (Stanovich and West 2008, Stanovich et al. 2013).

So, what can philosophers do about this? Our suggestion is we can use experimental tools and methods to help remove the conditions that make that kind of bias possible in the first place. To see how, let's start by thinking more generally about how tools work. They work by *transforming* tasks at hand, turning physical or epistemic tasks that are hard for us to perform into ones that are easier for us to perform. For example, screwdrivers transform the hard task of using your hand to embed a screw into the easier task of using that same hand to turn a shaft, and tape measures transform the hard task of estimating length by eyesight into the easier task of using eyesight to read what number is inscribed at some place on the tape. Let's focus on this second example. One particular way that tape measures help make it easier for us to measure things is by narrowing the amount of ambiguity in the signal that ultimately needs to be interpreted by us during the measurement process. Lots of epistemic tools work this way. Suppose, for example, that we are trying to uncover some chemical phenomenon regarding the acidity of a solution, and at first we only have the fairly crude method of using litmus paper to test pH. We are very likely to get stuck making judgment calls that allow for an undesirable degree of freedom, leaving it up to us whether a particular shade of pink is 6.4 or 6.5, for example. And in this kind of situation, it would not be surprising to find researchers supporting different ways of reading the results of these experiments, with researchers who support one hypothesis about the acidity of the solution reading the relevant test strips one way, and researchers who support alternative hypotheses reading them another way. Progress would halt, and researchers would be stuck in something we described in the previous chapter as an "immobilized stalemate." But we now have better ways of measuring the acidity of a solution, namely, pH-meters that provide unambiguous displays of a numerical result, and we can where necessary replace litmus paper with these kinds of meters. When the digital readout reads "6.5," then there is just no more wiggle room. Distinguishing "6.4" from "6.5" is an easy, unambiguous discrimination task for unaided human perception, which cannot be said of distinguishing two different though highly similar shades of pinkish red. Of course, there will be space for debating the reliability of that brand of meter, or on that sort of sample, and so on. But the more interpretive freedom that can be removed, the better shielded we are from ourselves as potential sources of error.

Our suggestion, then, is that experimental tools and methods can help philosophical inquiry by helping us remove ourselves as possible sources of interpretive ambiguity, which can interact in a methodologically toxic way with cognitive biases like myside bias. Rather than having to rely on our own estimates about how strongly held specific case verdicts are, or how central those particular case verdicts

may be, we can use standardized methods for gathering evidence and standardized measures of effect sizes. Social scientists currently use measures like Cohen's d, which is standardized against the baseline degree of variation in the data. But we think that philosophers should also develop more *philosophically meaningful* standard measures. For example, suppose we are considering how much particular structural change in a philosophical case involving stakes and knowledge attribution leads to a decrease in attributed knowledge; we could compare changing what's at stake with changes to the likelihood that the protagonist has misleading evidence.[17] Whatever experimental tools and methods that philosophers develop or adopt, it is worth pointing out that developing and adopting these kinds of experimental methods is perhaps one reason why science has *progressed* more easily than philosophy. Human cognition has remained less easily eliminated or avoided when generating evidence in philosophy than it has in science. Of course, the promise that experimental tools and methods have for helping philosophers do so does not rest so much in anything like the equivalent to reading pH-meters. After all, verdicts about philosophical cases measured using Likert scales and confidence ratings are quite different from readouts of a pH-meter. But that's okay because, in addition to the methodological pay-offs canvassed in this chapter already, experimental philosophy has yet other ways of paying off the investment that considerations of methodological rationality suggest are required here. As we will argue in the next chapter, one perhaps undersung, but very real, way that adopting experimental methods in philosophy can promote philosophical progress lies in the tools for analysis that come along with going quantitative, and the potential for transforming not just the observations themselves, but the inferences that we make on the basis of those observations; that is, the potential for helping us better establish better *phenomena* for philosophical inquiry.

[17] We discussed these kinds of cases in Chapter 3, and will again in the next chapter. See Weinberg (2014b) for additional discussion of philosophically meaningful measures of effect size, and Cova (2024) for an interesting application of calibration techniques to effect size measurement in experimental philosophy.

6
Experimental Philosophy and Philosophical Progress

The previous chapter laid out our case for substantially expanding our discipline's investment in experimental philosophy, because a failure to do so would be methodologically irrational: there are large, long-term methodological benefits to be gained from such an investment that would very likely far exceed the methodological costs. As described in that chapter, though, the benefits are described in rather abstract terms, as offering yet-unspecified instances of detection, correction, and management of errors. While such abstraction is surely no sin in a philosophy book, we nonetheless intend our work here to have practical implications for how philosophy gets done, and we hope for it to be engaged with on such terms. We thus think it would be useful for us to devote this chapter to laying out both more concretely, and in more detail, just what some of these methodological benefits might be, to help the reader envision what this all might look like in practice.

The first and most obvious concrete benefit of a wide adoption of X-phi is being able to determine for a large number of verdicts whether they are, in fact, subject to the vagaries that X-phi has indicated may be lurking in the verdict evidence on the whole. As we discussed in Chapter 2, we don't think that philosophers should expect anything close to a majority of cases that get considered by philosophers will prove problematic. The experimental challenge warns that, until and unless philosophers do *some* X-phi on them, we can't really tell in advance which ones will or won't be susceptible to unexpected variation, instability, or inconclusiveness. Yet that challenge can be taken and positively met—not from the armchair, of course, but from yet further experimental work. For example, Knobe (2021) has an excellent summary of the body of work that indicates that, *contra* some prominent earlier studies that had significant uptake from many foes of the armchair (including the authors of this book), there does not in fact seem to be any meaningful effects of incidental disgust on moral judgments (pp. 50–53). The more time and energy that gets invested in such work across the board, the more verdicts will be, as it were, cleared of charges, and be declared fit to use as data in our philosophical researches. Of course, some cases will be found to be beyond repair, at least for their originally intended purposes; for example, as noted earlier, both fake barn cases and Truetemp cases seem to consistently find modest majority rates of *positive* knowledge attribution, and thus we should at least for now take a hiatus on treating those cases as supporting theories that would deny they are knowledge.

Standing Up for Philosophy. Jonathan M. Weinberg and Joshua Alexander, Oxford University Press.
 DOI: 10.1093/9780191947704.003.0006

But this, too, is a kind of progress, in firmly removing as misleading evidence verdicts that previously had been relied upon.

This sort of separating-the-wheat-from-the-chaff methodological benefit is perhaps the easiest one for us to sell, since it falls fairly directly out of the nature of the experimentalist's challenge in the first place. We would clearly benefit from knowing which verdicts are robust, and which remain dubious. Nonetheless, we do not think that the benefits of investing in X-phi would stop there, and we will hereupon argue for a range of ways in which X-phi can help enable better research progress in philosophy. We will start by looking at recent arguments that contend that analytic philosophy has in fact made little progress over the course of its history, especially when compared to the empirical sciences and mathematics, and one popular diagnosis that the problem is that philosophers rely too much on arguments, and that argumentation turns out not to be a very good method for reaching consensus (Chalmers 2015, Daly 2017). We will argue, in response, that the problem has less to do with argumentation per se and more to do with a lack of solid *philosophical starting points*. We will recruit the notion that Jim Bogen and James Woodward (1988) describe as *phenomena* in their classic paper, "Saving the phenomena." Phenomena are the sort of enduring, theory-independent facts about the world that exist at an intermediate level between the grand sweep of theory and the messy nitty-gritty of observations.[1] Sciences can steadily accumulate established phenomena even as big-picture theories come and go, and constitute an important form of steadily growing scientific progress. One of the reasons that we think that analytic philosophy has not made as much progress as the empirical sciences and mathematics is that it has not been able to crystallize enough relevant *philosophical* phenomena, with philosophers too often looking to test their theories directly against their "observations," that is, against their own case verdicts, even though these are problematically messy. More importantly, we will argue that the tools of experimental philosophy are just the kinds of tools that can be used to help analytic philosophers establish more relevant philosophical starting points; not simply by providing better tools for observation, but more importantly, tools that help us to move from data to the kind of philosophical phenomena that can serve as stable philosophical starting points. To do this we will look carefully at several longstanding debates in analytic philosophy where we think that experimental philosophy has already helped us make substantial philosophical progress of the sort we will have in mind, and use these examples to highlight different ways that experimental philosophy can move the ball down the field.

[1] Some philosophers of science might worry about equating phenomena and "facts." The exact ontology of phenomena is up for debate, and they are sometimes thought of variously as patterns, regularities, abstract objects, ideal types, and so on. We have in mind something like *facts about the world* but nothing that we say here should be interpreted as our taking a stand on this debate. For further discussion in the philosophy of science, see Brown (1994), Leonelli (2009), Woodward (2011), Teller (2010), Kaiser and Krickel (2017), Colaço (2020). Thanks to David Colaço for helping us see this point.

1 Philosophical Progress

In recent years, there has been a surge of interest in whether and to what extent there is progress in philosophy. Opinions vary, all the way from one extreme that philosophy has made no progress whatsoever (Dietrich 2011) to the other extreme that philosophy has in fact answered all of its "big questions" (Cappelen 2017). Here we want to focus on a kind of *moderate pessimism* about philosophical progress that emerges when we compare philosophy to the empirical sciences and mathematics, where it seems that philosophy has made less progress than we see in those areas, especially when we measure success by the extent to which research communities are able to provide *consensus answers* to their central questions (Chalmers 2015).[2] Looking closely at active literatures, or even just chatting with a small group of philosophers who work on the same questions, one can easily get the sense that positive philosophical agreement is a rarity. But a further, more empirical basis for such moderate pessimism comes from a study conducted by David Bourget and David Chalmers (2014), and repeated more recently in 2020, where they surveyed English-speaking philosophers from around the world (though for obvious reasons, philosophers with Anglophone nationalities comprise the large majority of the sample). They found that even the most widely held views on a wide range of philosophical questions are not held that widely. These questions include both "classic" philosophical questions (What makes some actions right and others wrong? Is free will compatible with causal determinism?) and philosophical questions that have come to more recent prominence in contemporary philosophical discussion (What is the nature of mental content? What is the nature of perceptual judgment?) While it's surely wrong to operationalize "consensus" to mean anything even close to perfect agreement among practitioners, nonetheless their survey found almost no consensus whatsoever on these philosophical questions, even if we choose to operationalize it to mean as little as 80% agreement among practitioners. The only candidates for consensus answers on even this highly permissive way of operationalizing consensus are that first-trimester abortion is morally permissible (81.7% agreement) and that adult human beings have consciousness (95% agreement). Dropping the threshold for consensus even further to 70% still leaves out almost all of the philosophical questions that have come

[2] Even some philosophers who are more optimistic about philosophical progress, like Frances (2017) and Stoljar (2017), concede that philosophy does not lead to consensus answers to many of its central questions. Having said that, there is some disagreement about whether consensus is necessary for progress (Cappelen 2017 and Bengson et al. 2022), and more generally about how best to think about progress in philosophy (Dellsen et al. 2021). For a nice overview of several different ways of thinking about philosophical progress, including a discussion of the relationship between philosophical progress and the role of case verdicts in philosophy, see Gutting (2016). The fact that we are focusing here on consensus should not be read as dismissive of these other notions, but only as the selection of one legitimate notion of progress upon which to focus our discussion—a selection that was influenced by the fact that the sort of philosophical inquiry that uses the method of cases sure looks like the kind of inquiry that is *aiming* to converge on answers.

to define philosophy and philosophical inquiry, except for a bare handful of philosophical questions: Is capital punishment morally permissible? 75.1% of philosophers think that it is not. Is scientific realism true? 72.4% of philosophers think that it is. Would you enter Nozick's experience machine? 76.9% of philosophers say "no." Do we have knowledge of the external world? 79.5% of philosophers think that we do. Is *a priori* knowledge possible? 72.8% of philosophers think that it is. We can contrast these not exactly impressive percentages with the decidedly mixed results for the existence of abstract objects, whether internalism/externalism is true in epistemology, whether externalism is true about mental content, whether metaphilosophical naturalism is true, and many others. One particularly telling result for our purposes here is that only 56% of philosophers think that we have thus far acquired "a lot" of philosophical knowledge. Opinions in philosophy thus seem to vary even about how much opinions in philosophy may be expected to vary.

We concur with the moderate pessimists that, in terms of established consensus at least, philosophical progress has been, at best, underwhelming. So, why hasn't philosophy shown more progress? That is, what has made it so hard for philosophers to converge on answers to their questions? According to Chalmers (2015), much of the problem has to do with the fact that philosophers' most basic tool is argumentation, and argumentation turns out not to be a very good method for reaching consensus, because of a feature of argumentation he calls *premise deniability*:

> For most practitioners of philosophy, the phenomenon of premise deniability is familiar from both sides. As the old saying goes: one person's modus ponens is another person's modus tollens. When we give arguments for our views, we are frustrated to find opponents biting the bullet by rejecting what we took to be a plausible premise, without this serving as any sign of defeat. When we address arguments against our views, we sometimes work backwards from our rejection of the conclusion to see which premises we have to deny, and we deny them. In the best cases, we learn something from this, and we take on commitments that we might have antecedently found surprising. But these commitments are rarely untenable to maintain . . . As a result, philosophical arguments lead not to agreement but to sophisticated disagreement. (p. 18)

> There is less convergence in philosophy because the philosophical method has less power to compel agreement, and it has less power because of the phenomenon of premise deniability: arguments for strong conclusions in philosophy (unlike science and mathematics) almost always have premises or inferences that can be rejected without too much cost. (p. 25)

Of course, it is natural to ask why "premise deniability" is a greater problem in philosophy than the empirical sciences and mathematics. Here, Christopher Daly (2017) sheds some light:

> Still, even if argument is paramount in philosophy, why doesn't the arguing ever come to an end? In philosophy, our claims outrun our evidence in two respects. First, even where we agree about the evidence, it is not apparent which claim the evidence provides the most support for. Where the evidence is rich in philosophy, it tends to be disparate and conflicting, and thereby hard to assess. And where the evidence is meager, it provides little support for one philosophical claim over another. Second, in philosophy we often do not agree about the evidence. New evidence is always coming in just because new arguments are always being thought up. In debating about the new arguments we are debating whether they do provide evidence. And in the case of data besides argument, there is disagreement about whether such data as intuitions or phenomenology or parsimony principles are evidence or whether they are fundamental or whether they provide much support (see van Inwagen 2004, p. 335 n.4). Like a fractal, with every inferential step and with every appeal to evidence in philosophy, debate and argument can arise. (p. 37)

While this point could itself be debated—at the pain of self-refutation, how could it not?—we will nonetheless start from here: there has certainly been some progress in philosophy, but philosophy does not seem to progress anywhere near as steadily or significantly as the empirical sciences and mathematics. What we want to do is to juxtapose the diagnosis just rehearsed with a somewhat speculative observation about where philosophy has seen more or less progress. What seems to distinguish the more progressive parts of philosophy from those areas of philosophical inquiry that have seen less progress? It seems to us that one interesting possibility to explore is that the more progressive areas of philosophical inquiry are those that gain lots of useful traction by jostling up against other disciplines and practices that are themselves more progressive. For example, if we distinguish the more metaphysical parts of contemporary philosophy of mind from the parts that are better understood as the philosophy of psychology or cognitive science—say, questions about the modularity of perception, or mechanistic explanation in neuroscience, or the structures of our conceptual capacities—we see fairly steady progress in the more scientific parts of philosophy of mind, even if that progress has not been remotely total or totally monotonic. This doesn't mean, of course, that there has been no progress in the more metaphysical parts of contemporary philosophy of mind. The move from supervenience to grounding, for example, could perhaps be seen as a kind of philosophical progress.[3] Our point is comparative. We think that within philosophy itself there has been *less* progress in these areas than in areas of philosophical inquiry where philosophers engage meaningfully with members of

[3] Though probably not a consensus one; see, e.g., Wilson (2014), Turner (2016).

other research communities, and surely less than philosophers like Chalmers and Daly would like to have seen.

Let's just ride this conjecture for a minute. Assuming it is right, what could it tell us about the nature of this progress? The first thing to note is that even in the more progressive areas of philosophical inquiry, where philosophers engage meaningfully with people from different research communities, we don't see the kind of "big-question-answering" progress that Chalmers seems to have in mind. The grand debates about modularity, the structure of concepts, or the nature of consciousness are all still very much ongoing. So what makes these areas of philosophical inquiry progressive, assuming they are, cannot be measured simply in terms of whether or not philosophers working in those areas tend to agree more with one another about the answers to the largest questions that serve to organize research in those areas.

One might hypothesize that the greater progressiveness in philosophical areas on disciplinary borders could be explained in terms of their having, as it were, more arrows in their quiver. Bringing people from different research communities into contact with one another ought to yield a greater range of new modes of argument and a larger body of potential premises to draw from. But on further thought, this hypothesis can't be quite right. After all, there is plenty of *intra*disciplinary contact in philosophy, say, between ethicists and philosophers of language, and while such zones of exchange have been enormously fruitful in expanding the set of ideas and arguments across subdisciplinary lines, it does not seem to have led to any increase in philosophical convergence. In fact, having more argumentative resources turns out to be at best a mixed bag when it comes to progress, in philosophy or any other argumentative discipline, because additional argumentative resources can serve either to support existing arguments *or neutralize them*. The more wide open any field's argumentative repertoire is, the greater chance that increased argumentative resources will inhibit, not aid, its progress. The core of Chalmers' diagnosis for the comparative lack of progress in philosophy, after all, is not that philosophy lacks plausible premises, or varied and interesting ways to argue from them, but instead that it lacks the right kinds of premises that cannot be denied without incurring significant dialectical costs. And our current stockpile of argumentative armaments includes far too many ways of blasting away at an unwanted premise in an opponent's argument, with far too few ways of defending one's own premises adequately against counter-battery action.

Philosophers seem to use something like the following methodological norm in practice: a premise *p* can be denied by philosopher *S* provided that *S* can offer an argument against *p* that is *prima facie* valid and has plausible premises, even if those premises are themselves also plausibly deniable.[4] Put another way, philosophers

[4] Note this is trivially the case when not-*p* is itself antecedently plausible.

can plausibly deny any premise they want so long as they can provide reasons for doing so—but those reasons need not be either compelling or decisive, but merely plausible in their own right. Plausible deniability is, thus, too cheap in philosophy, and can be bought in its own currency. And this is at least one of the reasons why philosophy has so many spectacularly unmoveable stalemates.

While premise deniability can take a great many forms—that's part of why it is so enervating to philosophical progress—two especially widespread and potent modes of deniability operate openly under a banner upon which is written the words: "That's an empirical question." A hugely deleterious consequence of our entrenched commitment to armchair methods is that we have developed a norm under which empirical questions are ruled out of bounds for further philosophical inquiry, under the very weak plausibility norms just discussed. One denial mode that parasitically infests armchair methods, then, is the most straightforward: when one philosopher deploys the method of cases to render a case verdict, another can simply assert that their own verdict about the case diverges from the one reported—maybe they would render the contrary verdict or maybe they find their verdictive capacities nonplussed by the case. While philosophers can of course suspect the sincerity of such denials, recall our discussion from Chapter 1 about the many psychological vagaries that afflict the case verdict evidence. Philosophers' divergent verdicts may be due to such factors as an unanticipated sensitivity to variable aspects of the presentation of the case, or the circumstances under which they are considering it; real demographic variation in who renders which verdicts, whether systematic or unsystematic; or perhaps motivated cognition unconsciously turning a slightly warm but inconclusive verdict into a hot and bright one. Trapped in the armchair, without any way to explore such hypotheses, or indeed without any way to take a legitimately quantitative perspective, such divergences turn into a sort of unwinnable game of *p*-said/not-*p*-said.[5]

More complicatedly, but usually much more interestingly, a philosopher wishing to exercise premise denial on a case verdict can offer a conjecture as to how that verdict may in fact have been arrived at via some epistemically flawed psychological process. Here's the example that Ichikawa (2009, 94–95) uses to lead off his discussion of this sort of argumentative maneuver:[6]

[5] This marks another way in which our way of thinking about experimental philosophy is different from the way that Joshua Knobe thinks about it, especially in his work that we discussed in Chapter 2. He tends to default to thinking that everything is signal, so that when experimental philosophers find that people have both an inclination toward *p* and yet also an inclination toward not-*p*, Knobe takes this to mean something deep to be attended to and theorized. We tend to default to thinking that this means that there's some substantial noise here that needs to be filtered out. Such meta-level disagreements as that between us and him are themselves to be dispelled by the establishment of relevant phenomena.

[6] We have not much addressed the armchair appeal to verdicts about *principles*, instead of generalizations, both because their status in our practices is more contested, and because, frankly, there has been very little X-phi on them. But everything about the Sosa example generalizes to the rules of explaining away case verdicts, *mutatis mutandis*.

> In his "How to defeat opposition to Moore", Ernest Sosa (1999) suggests that safety is a necessary condition for knowledge.
>
> *Safety* S knows that *p* only if, S would believe that *p* only if *p* were the case.
>
> And he argues that, the "undeniable intuitive attractiveness" (Sosa 1999, 141) of such a necessary condition notwithstanding, sensitivity is not necessary for knowledge:
>
> *Sensitivity* S knows that *p* only if, were *p* not the case, S would not believe that *p*.
>
> Sosa apparently feels some pressure towards explaining away the intuitive appeal of Sensitivity; he attempts to do so by suggesting that safety and sensitivity, being counterfactual contrapositives of one another, are not equivalent, but might easily be thought to be. When one embraces Sensitivity, one is reacting to the truth of Safety, and, confusing the one for the other, one announces that Sensitivity is true.

What we particularly want to draw attention to, however, is how Ichikawa goes on to evaluate Sosa's denial move here. Ichikawa acknowledges that Sosa is committed to a "psychological claim" (p. 96) about how people may be prone to confusion about the contrapositives of various counterfactuals, but he does not think Sosa has done anything amiss by not seeking out substantive empirical confirmation of that claim. While wishing Sosa had illustrated the claim with a few more examples, he goes on to evaluate as follows:

> Someone more ambitious might cite more explicit psychological data that directly established such patterns. But I am inclined to think that Sosa is on firm ground in suggesting that many people do not ordinarily distinguish counterfactual conditionals from their contrapositives. Indeed, many people might, after brief reflection, endorse the equivalence of contrapositive counterfactuals. It takes a bit of cleverness to come up with a counterexample to the equivalence [of contraposed counterfactuals]. (p. 96).

We would note that, on Ichikawa's telling, any such "ambitious" appeal to actual psychological science is methodologically supererogatory; this is meant to be a technique fit for use in the armchair. But this also brings us to a deeper problem revealed in Ichikawa's discussion—or rather, what is revealed in what he does *not* discuss. He summarizes his three conditions on a successful "explaining away" thus:

> The case studies suggest that to explain away an intuition, one offers a psychological thesis to explain why people would have the offending intuition, even if it were not true. One's explaining away is successful insofar as the psychological

> thesis (a) is true, (b) predicts the offending intuition, and (c) does not depend upon the truth of the offending intuition. (p. 107)

What strikes us is how much weaker that second clause is than what seems to us it should be: that the psychological factor being posited is *actually involved* in producing the problematic intuition. Many hypotheses about underlying psychological mechanisms will predict the observed verdict, without actually being at play in the production of it. For instance, suppose that we grant that people really are capable of just the sort of modal confusion that Sosa postulates. That psychological fact would not tell us whether this confusion is, as a matter of fact, truly responsible for the reported attraction to sensitivity as a condition for knowledge. It may well be that, when we contemplate the Sensitivity counterfactual, we experience no such confusion, and really do find it intellectually attractive in its own right. Then Sosa's hypothesized psychological tendency, even if true as a generalization about human cognition, would have no power to disable a sensitivity premise. (We will just grant for the sake of argument that Sensitivity is substantially intuitive in the way Sosa claims.)[7]

It is unsurprising, though, that such a stronger condition of actual involvement is not on offer in an account of the relevant armchair practices, because those sorts of specific causal claims fall patently outside the range of our armchair capacities, for all of the reasons that we discussed in Chapters 1 and 5. And that very armchair inadequacy is part of why "explaining away" is so often an inconclusive, gridlock-making dialectical move. There is experimental evidence, after all, that when people learn about some general psychological foible, they tend to be much quicker to attribute those foibles as active in other people's cognition than to find it in their own, especially in conditions favorable for motivated cognition.[8] So it may be unsurprising if, say, a pro-Sensitivity epistemologist were to find themselves sincerely unmoved by Sosa's proposed diagnosis.

Ichikawa (2010) makes a solid case that explaining away undesirable verdicts is far from trivial, and he points to several instances of failed attempts. Nonetheless, this counter-premise dialectical technique seems to be too easy to deploy, too easy to make merely plausible conjectures about underlying psychological mechanisms. That's the danger: when there are too many counter-premise moves that are available, but which are only held to the standard of empirical plausibility, it will

[7] Consider results such as Goldin and Rouse (2001) on how "screened" auditions for orchestras seems to have led to a marked increase in the number of female musicians hired, and how they are so helpful in really demonstrating the existence of sexism in the audition process itself. When hiring committees hear auditions without being able to see who is performing, they turn out to hire considerably more women than in unscreened conditions. Without such results, it remained only a conjecture—albeit a highly plausible one—that sexism in the auditioning process was significantly to blame, as opposed to any actual sex disparities in ability, or, say, sexism in the training and recruitment process prior to the auditions themselves. But such empirical work yields the hard-to-deny premise that sexism in the auditioning process itself was at least one major culprit actually playing a causal role.

[8] For example, Babcock and Loewenstein (1997).

be extremely hard to secure the needed sorts of undeniability. (One might say that Ichikawa's principle leads us into a soft skepticism of low expectations.) What we want to suggest, then, is that what makes some areas of philosophy more progressive than others is not that there are *more* argumentative moves available in the more progressive communities. It's not like there has ever been a lack of clever moves to be made in philosophy, after all. What's been missing, especially in less progressive areas of philosophical inquiry, has been some way of figuring out how to keep clever philosophers from deploying that very cleverness to wriggle out from underneath the argumentative weight of good philosophical arguments. Instead, the more progressive zones within philosophy have had access to the kinds of resources, in the form of hard-to-deny premises, that *preclude* certain kinds of argumentative moves. In short: consensus *conclusions* are hard to arrive at in the absence of consensus *premises*.[9] We conjecture that the reason borderland zones of philosophy may be more progressive is that they can more easily import consensus results from the extra-philosophical fields with which they conduct their intellectual traffic. (This also explains why intra-philosophical trade doesn't bring the same benefits. If neither epistemology nor ethics, say, have adequate consensus premises of their own, then they cannot be of much help to other subfields in supplying such premises.)

We would note that one key methodological component here is that the philosophers respect those consensus results from the neighboring disciplines, *not* treating them as too easily challengeable on the basis of armchair plausibility judgments. This seems to us generally to be the case: philosophers of *x* tend to give high, though certainly not unquestioning, deference to the first-order *x* researchers. For example, Kourken Michaelian, Dorothea Debus, and Denis Perrin edited a recent (2018) *New Directions in the Philosophy of Memory*, and in their editorial introduction they affirm that,

> there is now a stable consensus on the active, reconstructive character of remembering. Whereas philosophers of memory until recently treated this as something to be defended (or attacked) by means of argument, the claim that memory has a reconstructive character now most often serves as a starting point for further argument. The psychology of memory is univocal in its endorsement of a reconstructive view, and this substantive consensus is no doubt in part a product of the methodological consensus [of relying on empirical psychological findings]. (p. 6)

This methodological recruitment from allied fields does not always have to manifest as an appeal to science, however. In the philosophical aesthetics, any proposed theory of art must take very seriously such strange artworks as Marcel Duchamp's *Fountain*, where he took a manufactured men's porcelain urinal, turned it upside

[9] Consensus on acceptable modes of inference is also important, of course, but we won't pursue this here.

down, and signed it with the fake name "R. Mutt"; or John Cage's infamous composition *4'33"*, in which the assembled musicians play exactly no notes whatsoever for an interval of that length. One could easily imagine philosophers contesting the artwork status of such "creations" on all sorts of plausible grounds, yet this is by and large just not done.[10] And the reason we do not find any such contestings—any such attempted premise-denials—is because the "artworld" of artists, critics, and art historians have long since arrived at a positive consensus about their status, and that consensus is then deferred to by the aesthetics community.[11]

We'll now leave aside our conjecture about which parts of philosophy may be more or less generally progressive than others, as a topic for further empirical investigation, and focus on a question suggested by our discussion so far. Namely, if some parts of philosophy can get consensus premises by reaching for them from outside of philosophy proper, what can other subdisciplines do within philosophy's own borders to generate such premises for itself—a sort of domestic import substitution? Our proposal here is that, with the help of experimental philosophy, we can indeed establish our own *philosophical phenomena*.

2 Philosophical Phenomena

By invoking the term "phenomena" here, we have in mind something like what Jim Bogen and James Woodward have in mind in their famous discussion of phenomena in science (Bogen and Woodward 1988; Woodward 1989, 2000, 2011). According to Bogen and Woodward, it is a mistake to think of scientific theories as directly facing a tribunal of data, in the sense of a body of individual observations. Data are much too messy to serve that role. Rather, there is an intermediate level of concrete, more or less theory-independent established claims about the world that do this work: *phenomena* that are uncovered through a process of carefully sifting through and systematizing that mess of observations, and that can thereupon provide the right sort of *explananda* to inform and constrain scientific theorizing. To illustrate the distinction, they write (1988, 306):

> Examples of data include bubble chamber photographs, patterns of discharge in electronic particle detectors and records of reaction times and error rates in

[10] Though there are a handful of dissenters, they are mostly either not recent (e.g., Beardsley 1982, Cohen 1972) or generally seen, even by themselves, as taking a radical view (e.g., Zangwill 2002). As Lopes (2008) observes, the status of the expert verdicts on such cases as *art* is not really disputed ("Nobody attending a performance of *4'33"* doubts that they are in the presence of art" (p. 121)), though what perhaps can be disputed is whether, e.g., Cage's "performance" can properly be categorized as *music*, per se. Our thanks to Nat Hansen and Aaron Meskin for discussion on this point.

[11] See Weinberg (2018) for further discussion of how case verdicts in aesthetics may be in better methodological shape than other philosophical subdisciplines, in large part because of its access to consensus verdicts from expert communities.

> various psychological experiments. Examples of phenomena, for which the above data might provide evidence, include weak neutral currents, the decay of the proton, and chunking and recency effects in human memory.

Here are some passages from one of Woodward's (1989) classic papers to further illustrate the distinction, and to begin to sketch some conceptual connections to our work in preceding chapters on such matters as noisy or biased sources of evidence—connections we will further draw out later:

> Phenomena . . . are relatively stable and general features of the world which are potential objects of explanation and prediction by general theory. Examples of real or putative phenomena . . . include weak neutral currents, gravitational radiation, Brownian motion, proton decay, capacity limitations and recency effects in short term memory, and the proportionately higher rate of technical innovation among middle-sized firms in moderately concentrated industries. Data, by contrast, play the role of evidence for claims about phenomena. (p. 393)

> There is another way of thinking about the distinction between data and phenomena which many scientists seem to find natural and suggestive . . . scientific investigation is typically carried on in a noisy environment; an environment in which the data we confront reflect the operation of many different causal factors, a number of which are due to local, idiosyncratic features of the instruments we employ (including our senses) or the particular background situation in which we find ourselves. The problem of detecting a phenomenon is the problem of detecting signal in this sea of noise, of identifying a relatively stable and invariant pattern of some simplicity and generality with recurrent features—a pattern which is not an artifact of the particular detection techniques we employ or the local environment in which we operate. Problems of experimental design, of controlling for bias or error, of selecting appropriate techniques for measurement of data analysis are, in effect, problems of tuning, of learning how to separate signal from noise in a reliable way. (pp. 396–397)

Once they emerge in the course of inquiry, then, phenomena are excellent candidates for the kind of *undeniable*, or at least *highly costly to deny*, fixed points that seem to be so badly lacking in much philosophical inquiry. Cast in the terms we just introduced, phenomena are excellent candidates to serve as the kind of *consensus premises* that Chalmers diagnoses as missing from long standing philosophical debates. They, not the observations, are well suited to constrain and inform productive theorizing. And it seems that we do find that the comparatively more progressive corners of philosophy tend to be the ones that have a greater wealth of phenomena to draw upon. Philosophies of the sciences gain phenomena as their target sciences do, which is certainly one of the reasons why work in philosophy of psychology and

cognitive science has progressed so rapidly in recent years. To take just one from our earlier list of examples, Ophelia Deroy (2014) provides a wonderful survey on the role that empirical work has played in discussions about the modularity of perception. Among the work that she highlights is work on cognitive penetrability (see, e.g., Hansen and Gegenfurtner, 2006, Hansen et al. 2006, and discussion by Siegel 2011, Macpherson 2012), multisensory perception (for discussion, see O'Callaghan 2011), and synesthesia (for discussion, see Segal 1997). But this is not the only place in philosophy where we see progress being made when philosophers connect to what's going on in other disciplines. Philosophy of language is quite obviously informed by phenomena from linguistics, and as noted above aesthetics is able to draw on well-entrenched artworld practices of creation and criticism, as well as on relevant work in psychology and the social sciences. But what about epistemology, or metaphysics, or ethics? What are the relevant phenomena, and where and how do we find them? The answers to these questions are not quite so obvious, although the consensus in analytic philosophy seems to be that case verdicts provide us with some *access* to the relevant philosophical phenomena. As we discussed in Chapter 1, after all, this is what the method of cases is supposed to be all about: testing our epistemological or metaphysical or ethical theories against what we think about real or imagined cases. What we want to suggest in the next few sections is that experimental philosophy can help philosophers *better understand* the relevant philosophical phenomena that the method of cases was meant to help us uncover in the first place.

All of this talk about philosophical phenomena raises an important metaphilosophical question that we've managed carefully to avoid having to address until now.[12] That question is whether to accept what Alvin Goldman and Joel Pust (1998) call a *mentalist* conception of philosophical theories, according to which they tell us something about how people *think* about things like knowledge, moral responsibility, and moral permissibility, to return to the examples we used in Chapter 1, or an *extramentalist* conception, according to which philosophical theories tell us something about things like knowledge, moral responsibility, and moral permissibility themselves. We are going to punt on this question, for a couple of reasons. The first is somewhat personal: during the course of writing this book we found that we disagree about the answer to this question! (For those who are keeping score at home, Josh is a mentalist and Jonathan is an extramentalist. And each of us is disappointed in the other person's poor philosophical judgment in this regard.) The second reason for punting on this question is perhaps more principled, and takes us back to both the philosophical folk wisdom about the method of cases that we discussed in Chapters 1 and 5 and the issues about widespread philosophical dissensus from earlier in this chapter. For this question, even though it concerns a fundamental metaphilosophical matter, seems to us to be yet

[12] Our thanks to Sam Bennett, Ron Mallon, Aaron Meskin, Hannah Rakoski, and Rissa Willis for pressing us on this point at the Athens Philosophy Workshop.

one more of those questions that simply has not produced a consensus answer in the philosophical community at large. And since the philosophical folk wisdom that surrounds the armchair method of cases presupposes no specific answer to this question one way or the other, there's no reason why our arguments against that method would need to take on a contentious answer to this question.

There is another reason for wanting to avoid having to answer this question: we think an easy waffle is available to us here in the form of a constructive dilemma. To philosophers who support a mentalist conception of philosophical theories, we can simply point out that it seems relatively straightforward that experimental philosophy should be able to help us better understand the relevant philosophical phenomena, since on the mentalist view those phenomena *just are* psychological phenomena. And most of the rest of this chapter is meant to illustrate a number of different ways that experimental philosophy can help us better understand the relevant philosophical phenomena understood in this way. To those philosophers who support some kind of extramentalism about philosophical theories, we can ask them to tell us first what *they* think is the link between case verdicts, on the one hand, and philosophical facts, on the other, and we can just plug that into the story that we want to tell in this chapter. Verdict-deploying philosophers who support some kind of extramentalism about philosophical theories must endorse some version of what we called in Chapter 3 a moderate reliability thesis about the case verdict evidence, and so are committed to *some* sort of connection between our rendering the verdict that *p* in the method of cases and *p*'s being true as a matter of philosophical fact.[13] Let's call this a *vindicating theory* for the case verdict evidence. Not all vindicating theories are equivalent. For example, different vindicating theories may involve subtly different views about how to apply the competence/performance distinction to case verdict evidence. As we argue in our (2010) along with Ron Mallon, there is no easy way to infer to a philosophical competence/performance distinction from a psychological one—it is not something that can be just read off the science, even for philosophers who are sympathetic to mentalist metaphilosophical views. Nonetheless, we will presuppose that the range of plausible vindicating theories will overlap substantially, and in a way that will easily cover the examples that we will focus on for the rest of this chapter. Just to be clear, a verdict-deploying practitioner does not need to have any specific vindicating theory in mind. They just need to recognize their commitment to the truth of some-vindicating-theory-or-other.

[13] Philosophers who support some kind of extramentalism without endorsing at least a moderate reliability thesis about case verdict evidence would likely be stuck in some kind of philosophical skepticism, and we are not looking here to retail any positive reasons for optimism to sell to such poor souls. But what experimental philosophy can offer them is the dark, Macherian hope that experimental philosophy will help establish that, as it turns out, there just aren't many philosophical phenomena to be uncovered in the first place. All in all, though, this chapter is just not directed to such philosophers.

Long story short: even for the philosophers who support some kind of extramentalism, establishing psychological phenomena relevant to the case verdict data, or to any other philosophically laden aspects of our mental lives, will thereby serve as establishing correspondingly hard-to-deny philosophical phenomena, as well. Our discussion at the start of this chapter is relevant here: where an extensive and thorough empirical investigation reveals some particular case verdict to be robust, then even philosophers who support any kind of extramentalism should take the truth of that case verdict to have not just a *prima facie* positive status, but something much stronger: someone who would deny that case verdict would take on a special and very high burden to do so. The premise-evading strategy of straightup denial thus could become dialectically costly in a way that it was not when attempted from the armchair. In short, such robust case verdicts would count as the kind of philosophical phenomena that could thereby serve as the requisite sort of argument-stopping premises. Allowing ourselves this waffle, then, going forward we will just say "philosophical phenomena" and allow readers to fill in for themselves what sort of mentalist or extramentalist approach to such phenomena they would prefer to take.

Importantly, philosophical phenomena (on either mentalist or extramentalist construal) are not just limited to establishing robust case verdicts. One of the important lessons that philosophers can learn from the social sciences is that we should never put too much weight on any one study in isolation. What matters are enduring or recurring *patterns* observed across a substantial number of studies. At this point in epistemological inquiry, for example, it is less important to figure out whether this or that specific case is a Gettier case and much more important to figure out the general and robust patterns of unknowing justified true belief.[14] Of course this is not an alien idea to analytic philosophy. Debates about, say, the doctrine of double effect in ethics or stakes effects in epistemology (see below) are fundamentally about patterns even more so than they are about individual cases per se.

Robust individual case verdicts and robust verdict patterns can thus serve as philosophical phenomena in a fairly direct way, by constituting the sort of evidence that philosophical theories ought to be responsible to. What we have been calling "error profiles" is another important form of phenomena, operating perhaps at a meta-level. There are practices of premise-denial in science too: it's just that those practices are themselves tightly disciplined by empirical facts on the ground. When scientists present their findings, and another scientist raises the plausible, empirically live concern that what they are reporting is an artifact of their instrument, one way they might want to be able to reply is that we know that *that* instrument doesn't induce *that* sort of artifact—or that *these* are the steps to take to preclude such artifacts, and you indeed took such

[14] See Blouw et al. (2018) for discussion.

known precautionary measures. When our instruments are our own minds, then establishing phenomena about the *underlying mechanisms* of verdict-production can help to preclude the kinds of "explaining away" deadlocks that we discussed in the previous section: any proposed hypothesis as to the mechanism producing or influencing our verdictive reports can be tested, and ultimately confirmed or disconfirmed. We think that all these sorts of philosophical phenomena can sharply increase the cost of denying premises, and so can sometimes help philosophy out of the disputatious mire in which Chalmers and Daly have found us collectively stuck.

One key promise for experimental philosophy, then, that we will be exploring in the rest of the chapter is its capacity to move us closer to philosophical phenomena, by excavating attempted premise denials out of the boggy quagmires of armchair plausibility-mongering and transplanting them to the more green, fertile, firmer fields of empirical inquiry (if nonetheless messy in their own way). Our discussion in Section 1 about armchair philosophy's inability to make philosophical progress involves a worry about armchair philosophy's reliance on armchair psychology, and so immediately suggests the remedy: standing up, and using the tools and methods of experimental psychology and cognitive science to help settle these empirical disputes, rather than allowing them to fester just outside of our disciplinary bounds. It's not so much that experimental philosophy lets us play the game of premise-deniability better, so much as it can *radically change the rules of the game*. Instead of a norm of empirical plausibility countering empirical plausibility, which embed our inquiries into an ultimately boring version of the Seussian North-Going Zax and the South-Going Zax, attempts to explain away can become productively embroiled in *addressable* empirical matters. There is much methodological benefit to be had in transforming "that's an empirical question" from a philosophical discussion-stopper to a sign that further philosophical debate is to be had—only now, such debate both can and must be conducted from a standing position.

To get a sense for what we have in mind here, we can again look at the way that philosophers of psychology talk about the role that empirical work has played in the development of philosophical work on memory. Here again is Michaelian, this time writing with John Sutton in the *Stanford Encyclopedia of Philosophy* entry on memory (though we believe the trend is general across naturalistically minded parts of the profession):

> Much of the impetus for the emergence of the field was due to a trend, beginning in the late 1990s, towards increased interdisciplinarity among philosophers working on memory . . . a trend which reinvigorated and transformed older philosophical debates by bringing them into contact with empirical and theoretical developments in psychology and the sciences of memory more broadly.

We will demonstrate, over the course of the next several sections, that there are already prominent examples where experimental philosophy has helped us make significant philosophical progress by helping us move toward locking down the phenomena relevant to philosophical debates that were previously deadlocked. In each of these examples, experimental philosophy has helped move debates forward in precisely the way that Woodward and Bogen describe in the passages quoted above by helping philosophers "separate signal from noise in a reliable way."

3 Sorting Out Psychologically Intermediate Causes in Debates about Free Will

There has been a lot of philosophical work throughout the history of philosophy trying to understand the relationship between free will and causal determinism: are freedom and moral responsibility possible in a deterministic universe? One striking fact about this work is that it seems mired in what Dylan Murray and Eddy Nahmias (2014) call a "dialectical stalemate" very much like the kind of situations that Chalmers and Daly are worried about:

> Compatibilists typically agree that if free will were what incompatibilists say it is—a type of freedom that requires having an unconditional ability to do otherwise or being the "ultimate source" of one's actions—then it would be incompatible with determinism. Incompatibilists typically agree that if free will were what compatibilists say it is—a type of freedom that requires a less metaphysically demanding set of capacities, such as reflective, rational self-regulation of one's actions—then it would be compatible with determinism. Each side believes that the other is wrong about what free will is, and about what conditions are required for having it, but they agree on which conditions are compatible with determinism and on which are not. (pp. 434–435)

Can experimental philosophy help, and if so, how? One obvious suggestion is that experimental philosophy could help by getting us some data that would be relevant to the debate between compatibilists and incompatibilists. We could see, for example, what people think about cases involving free will, moral responsibility, and causal determinism. Interestingly, this kind of experimental work has turned out actually to exacerbate the stalemate somewhat, at least at first glance. Some early experimental work that focused on finding out what people think about these kinds of cases suggested that people think that free will and moral responsibility are *compatible* with causal determinism (for example, Nahmias et al. 2006), and some early experimental work on these kinds of cases suggested that people think that free will and moral responsibility are *incompatible* with causal determinism (for example, Nichols and Knobe 2007). The data, it turns out, are messy.

This would be a problem if philosophical theories aim primarily to explain some body of philosophical data; since the data point in different theoretical directions, it would seem that experimental philosophy provides little promise of breaking the dialectical stalemate. But, on the view that we are recommending here, philosophical theories, like scientific theories, aren't meant to explain philosophical data, even the kind of data that experimental philosophers can help us collect, but are instead meant to explain stable philosophical phenomena that need to be uncovered through a process of carefully locating the relevant philosophical signal through the experimental noise. And what's important for our purposes here is that experimental philosophy provides essential help in doing this kind of work.

To see how, let's follow the example a little bit further. Murray and Nahmias hypothesized that incompatibilist case verdicts might rest on a mistake, or rather on the difficulty that people seem to have grappling with some of the fundamental features of causal determinism (Nahmias et al. 2007, Nahmias and Murray 2011). In particular, they hypothesized that incompatibilist case verdicts rest on the mistaken belief that causal determinism involves some kind of *bypassing*, where our desires or intentions are bypassed completely in whatever causal story explains our actions. To test this hypothesis, Murray and Nahmias asked participants to consider versions of the cases that were central to the empirical studies reported by Nahmias et al. (2006) and Nichols and Knobe (2007), and then asked them a series of questions meant to gauge whether they judged that the protagonists described in the cases acted freely and whether their desires or intentions played a role in the causal story that explains those actions.[15] For example, they had participants consider the following two cases:

> Imagine a universe (Universe A) in which everything that happens is completely caused by whatever happened before it. This is true from the very beginning of the universe, so what happened in the beginning of the universe caused what happened next, and so on right up until the present. For example one day John decided to have French Fries at lunch. Like everything else, this decision was completely caused by what happened before it. So, if everything in this universe was exactly the same up until John made his decision, then it had to happen that John would decide to have French Fries.
>
> Now imagine a universe (Universe B) in which almost everything that happens is completely caused by whatever happened before it. The one exception is human decision making. For example, one day Mary decided to have French Fries at lunch. Since a person's decision in this universe is not completely caused

[15] Murray and Nahmias conducted three different studies. We are focusing here on the first study. The results of the different studies converged to support the hypothesis that people are more likely to give incompatibilist responses when they mistakenly believe that causal determinism involves some kind of bypassing.

> by what happened before it, even if everything in the universe was exactly the same up until Mary made her decision, it did not have to happen that Mary would decide to have French Fries. She could have decided to have something different.
>
> The key difference, then, is that in Universe A every decision is completely caused by what happened before the decision—given the past, each decision has to happen the way that it does. By contrast, in Universe B, decisions are not completely caused by the past, and each human decision does not have to happen the way that it does.

And, then asked participants to indicate their level of agreement with the following statements:[16]

> In Universe A, it is possible for a person to have free will.
> In Universe A, what a person wants has no effect on what they end up doing.

And they found that the more likely participants were to interpret cases involving causal determinism to be cases where a person's desires and intentions do not play a role in the causal story that explains those actions, the less likely they were to think that the person whose actions were described in those cases acted freely, and vice versa. In short, it seems that people are more likely to think that incompatibilism is true when they mistakenly think that causal determinism involves bypassing.

But the story isn't over yet. In fact, it has a twist. It turns out that people only think that causal determinism involves bypassing when the states being bypassed are mental states. People do not think that causal determinism bypasses physical states or processes, for example, the kinds of processes that cause a volcano to erupt, and they do not think that causal determinism bypasses artificial states or processes, for example, the kinds of states or processes involved in computer programs. So there's a puzzle here that hasn't yet been answered: why do people tend to think that causal determinism involves bypassing only when the states in question are mental states? David Rose and Shaun Nichols (2013) have a proposed an interesting answer to this question. They start with two possible causal models. According to the first model, the way that philosophers (sometimes) describe causal determinism causes people to mistakenly think that it involves bypassing, and this influences how people think about the relationship between freedom and determinism. According to the second model, the way that philosophers (sometimes) describe causal determinism causes people to think that freedom is incompatible with determinism, and this influences how people think about the role that desires and intentions play in the causal story that explains our actions. Murray

[16] These are just two of the statements that they used. Other statements focused on moral responsibility and blame, as well as causal roles played by other kinds of intentional states. Again, the results converged.

and Nahmias argue for the first model, and they use a mediation analysis to provide evidence for it. But, as Rose and Nichols point out, the kinds of regression analyses that are used when testing mediation are correlational, and as we know, correlations don't completely answer causal questions—in particular, correlations are symmetric in a way that causal claims are not. So, while the regression analysis that Murray and Nahmias use provides *some* evidence for the first model, it doesn't provide *conclusive* evidence for that model, and in particular, it is consistent with the rival model.

To try to figure out which model is correct, Rose and Nichols conducted another study using cases very much like the ones used by Murray and Nahmias, which were themselves very much like cases used in earlier experimental studies, as well as in more traditional philosophical debates between compatibilists and incompatibilists. The difference is that Rose and Nichols used structural equation modeling rather than a regression analysis to interpret the data. What they found is that the way that philosophers describe causal determinism causes people to think that freedom is incompatible with determinism, and this influences how people think about the role that desires and intentions play in the causal story that explains our actions. That is, the participants' bypassing judgments are caused by their incompatibilism, not vice versa. And this in turn helps explain why people think that causal determinism involves bypassing only when the states being bypassed are mental states.

What's the moral of the story? We started with an example of the kind of impasse that Chalmers is worried is characteristic of philosophical inquiry; what Murray and Nahmias describe as a "dialectical stalemate" between compatibilists and incompatibilists. And we've seen how experimental philosophy can help us move past this kind of impasse, not by providing additional data, but by helping philosophers separate signal from noise—that is, by helping us *move closer toward* the relevant philosophical phenomena. Recall Woodward's (1989) description of experimental design, when he wrote, "Problems of experimental design, of controlling for bias or error, of selecting appropriate techniques for measurement of data analysis are, in effect, problems of tuning, of learning how to separate signal from noise in a reliable way" (p. 397). The initial experimental data was noisy, with some early experimental work suggesting that people think that freedom is compatible with causal determinism and some early experimental work suggesting that people think that freedom is incompatible with causal determinism. But the debate doesn't get bogged down there. When mathematical inferential tools tune out the noise and locate the signal, we find a consistent phenomenon emerges in which people do think that freedom is incompatible with causal determinism. And this phenomenon can then be appealed to when debating our philosophical theories of freedom and moral responsibility.[17]

[17] It is important to emphasize that our claim is *not* that these studies have *resolved* the longstanding debates about freedom and determinism; remember that Bogen and Woodward's level of phenomena

One last note: the kinds of inferences that both sets of authors deploy here, in trying to dig down into the underlying mechanisms of folk incompatibilist responses, are fundamentally quantitative. There simply is no armchair version of correlational or structural equation modeling. One of the further great methodological benefits to be acquired by investing in X-phi, then, is the greatly expanded set of *inferential tools* that would come along as well.

4 Uncovering Artifacts and Discovering New Phenomena in Debates about Epistemic Pragmatic Encroachment

Here's another, and more controversial, example of how experimental philosophy can help establish what is or isn't a legitimate philosophical phenomenon, drawn from epistemology. This debate centers on a series of papers in which the authors argue that practical considerations, for example, the practical costs associated with being wrong, play such a significant role in knowledge attributions that they ought to deeply inform our theories of knowledge (Fantl and McGrath 2002, 2010, Hawthorne 2003, Stanley 2005, Hawthorne and Stanley 2008). Like many philosophical arguments, these arguments for *pragmatic encroachment* often involve the method of cases.[18] So, for example, much of this debate traces back to what people think about DeRose's (1992) famous "bank cases":

> My wife and I are driving home on a Friday afternoon. We plan to stop at the bank on the way home to deposit our paychecks. But as we drive past the bank, we notice that the lines inside are very long, as they often are on Friday afternoons. Although we generally like to deposit our paychecks as soon as possible, it is not especially important in this case that they be deposited right away, so I suggest that we drive straight home and deposit our paychecks on Saturday morning. My wife says, "Maybe the bank won't be open tomorrow. Lots of banks are closed on Saturdays." I reply, "No, I know it'll be open. I was just there two weeks ago on Saturday. It's open until noon."
>
> My wife and I drive past the bank on a Friday afternoon, as in Bank Case A, and notice the long lines. I again suggest that we deposit our paychecks on Saturday morning, explaining that I was at the bank on Saturday morning only two weeks

does not replace theory, but is an intermediary between data and theory. Rather, our claim is that future theorizing about free will needs to take on board as a costly point to deny that there is a significant incompatibilist element in our folk judgments in this area, and indeed the particular contours of that element.

[18] Here and elsewhere, there are other, less case-heavy arguments to be found for the various positions in the field. But, similar to how we have argued in Chapter 2, it would be false to the literature here to think that the method of cases has not played a huge evidentiary and dialectical role.

> ago and discovered that it was open until noon. But in this case, we have just written a very large and very important check. If our paychecks are not deposited into our checking account before Monday morning, the important check we wrote will bounce, leaving us in a *very* bad situation. And, of course, the bank is not open on Sunday. My wife reminds me of these facts. She then says, "Banks do change their hours. Do you know the bank will be open tomorrow?" Remaining as confident as I was before that the bank will be open then, still, I reply, "Well, no. I'd better go in and make sure." (p. 913)

While there is a large and nuanced literature trying to theorize what's going on with cases like these, aimed at explaining (or, sometimes, explaining away) just what it is that makes these two cases epistemologically different, everyone agrees that one key difference is how much more seems to be at stake in the second case than in the first case.[19] After all, DeRose makes a point to say in the second case that if they are wrong, they will be left "in a *very* bad situation." And, so much of the literature about pragmatic encroachment has agreed that there is at least a *prima facie* epistemic difference between low- and high-stakes cases, with knowledge attributions generally going down as stakes go up.[20]

Despite this strong local consensus that stakes affect knowledge verdicts one way or another, a growing body of X-phi literature appears to be converging on a different story. All of the early work used a paradigm that closely follows the armchair method, by confronting participants with vignettes and eliciting knowledge or not-knowledge verdicts. The results are a somewhat mixed bag of non-effects, or, at most, fairly small effects (see, e.g., Buckwalter 2010, Feltz and Zarpentine 2010, May et al. 2010, Sripada and Stanley 2012; see Dinges and Zakkou (2021) section 1 for a nice overview). This raises serious questions about the status of pragmatic encroachment as a candidate philosophical phenomenon.

At this point, it is important to mention that the noisiness of the early experimental data was intensified by concerns about how this early experimental work

[19] In addition to changing what is at stake for the protagonists, DeRose's second bank case also changes what specific possibilities of error have been made conversationally salient. This has given rise to a parallel debate in epistemology about the relationship between *salience* and knowledge attribution (see, e.g., DeRose 1992, Cohen 1999, Hawthorne 2003, Williamson 2005, Nagel 2010, Gerken 2013, 2017, and Turri 2017). For experimental contributions to this debate, see Schaffer and Knobe (2012), Hansen and Chemla (2013), Nagel et al. (2013), Alexander et al. (2014), Buckwalter (2014a), Buckwalter and Schaffer (2015), Turri (2015, 2017), Gerken and Beebe (2016), Waterman et al. (2018), Grindrod et al. (2019), and Gerken et al. (2020). This parallel debate has also generated a significant amount of experimental work and provides a nice example of another way that experimental work can help us make philosophical progress, namely, by helping to *affirm* and *refine* an epistemological theory. Interestingly, there is also experimental work that suggests that knowledge attributions are sensitive to more than what is at stake or what possibilities have been made salient in a given conversational context. See, e.g., Turri et al. (2016) on a broad range of "actionability judgments" and Jackson (2021) on "social context."

[20] We are frankly eliding here many significant distinctions in the literature, such as whether the relevant issue is what the stakes are for the agent under epistemic evaluation, or what they are for the epistemic evaluator.

was conducted, with worries being raised by philosophers on both sides of the debate about whether or not researchers were using the right kinds of cases, were making the right kinds of manipulations, and were accounting for potential confounds, including task demands and cognitive load (Buckwalter and Schaffer 2012, Schaffer and Knobe 2012, Sripada and Stanley 2012). But a batch of more recent studies (Francis et al. 2019, Rose et al. 2019, and Porter et al. 2024), using large international samples, have found yet more null findings for a meaningful influence of stakes on knowledge verdicts. The literature is all sufficiently recent and developing that one ought not to declare the debate settled.[21] Yet it would be too coy to say the trends so far are merely suggestively pessimistic for pragmatic encroachment theorists—for, at present and subject to possible future revision, there is no reason to accept the claim that there is a stakes effect on knowledge verdicts. This means that, if current trends continue, we will have a nice case of X-phi *debunking* a claim in the literature, one which has been agenda-defining for a significant part of epistemology, and upon which dozens and dozens of papers have depended. We thus would have an excellent demonstration of the idea we have been putting forward throughout this book, namely, that X-phi offers the potential to detect and correct for errors that the armchair cannot. Perhaps, as Gertrude Stein said about her childhood home in Oakland, there is no there there. It is important to note, as it was with Stein's observation, that those kinds of realizations constitute their own kind of progress. As Woodward (1989, 450) writes, "Learning to separate the real from the artifactual is an important kind of scientific progress, although it is not progress in the construction of explanatory theory."[22]

Such is the story to date for the standard sort of knowledge verdicts, that is, those in which we are presented with a hypothetical scenario and decide to apply or withhold an attribution of knowledge. But the experimental story on stakes effects has, intriguingly, gone beyond examining these kinds of verdicts, and has resulted in an interesting twist in the overall story of pragmatic effects in folk epistemology. In a recent set of studies reported in their (2019), Kathryn Francis, Philip Beaman, and Nat Hansen deployed two different kinds of experimental design. The first kind is the standard X-phi/armchair vignette design, which they call an "evidence-fixed" design, where participants were asked to think about a number of different high-stakes and low-stakes cases, and then were asked simply whether or not the protagonist in those cases knew the relevant claims. In our terms here, their participants offered verdicts about hypothetical cases. But they also used a

[21] For example, Turri et al. 2016 in particular seems to us an important outlier here. See also Dinges and Zakkou 2021, who do report something like a stakes effect on verdicts, but in a novel experimental task in which participants first make a positive verdict, are then confronted with information that the stakes in the scenario are high, and are thereupon asked whether they want to *retract* the initial verdict. We applaud this clever methodological innovation, but because it is so different from the traditional version of vignette methods, it is just unclear to us at this time how best to integrate this result with the overall trend of negative findings here.

[22] For interesting discussions of artifacts, see Feest (2022) and Craver and Dan-Cohen (2024).

second kind of experimental design called an "evidence-seeking" design, modeled on work pioneered by Ángel Pinillos (2012). Here participants were again asked to think about a number of different high-stakes and low-stakes cases, but were now asked *how much evidence the protagonists would need to collect* before they would count as knowing the relevant claims. Traditionally, the philosopher or experimenter stipulates in the vignette what the degree of evidence is, and then asks for a yay-or-nay knowledge verdict; here, participants are asked to adjust the degree of evidence to whatever point is needed for them to arrive at a stipulated positive knowledge verdict. So, for example, participants were asked to read something like the following vignette:

> Elaine is a medical researcher. Her task is to create a vaccine for a virus. Elaine has done this before, and she has a checklist that specifies all of the steps she needs to take to make the vaccine. Elaine is following all of the steps correctly. Elaine's assistant has informed her that there are 100 human research participants who have volunteered to trial the vaccine before it is distributed more widely. If Elaine does not follow the steps correctly, it will produce an ineffective combination that when administered to the research participants will kill them all after several days of excruciating pain.

In the "evidence-fixed" design, participants were then asked the extent to which they agreed or disagreed with the statement that "Elaine knows that she is making the vaccine correctly." By contrast, in the "evidence-seeking" design, participants were asked "How many times does Elaine need to consult her check list before she knows that she is making the vaccine correctly?" In keeping with the trend reported above, Francis and her colleagues did not find a stakes effect in the evidence-fixed design. But here's the twist: they *did* find stakes effects *in the evidence-seeking design*. This is consistent with Pinillos' original results, and has been replicated recently by Porter et al. 2024, so it has a very good, if still somewhat probationary, claim at being its own piece of epistemological phenomena.

We also find in this piece of the literature a nice demonstration of how X-phi can help steer clear of premise deniability quagmires. Wesley Buckwalter and Jonathan Schaffer (2015) argued, in seeking to reinterpret the earlier Pinillos findings, that the reason why stakes influence judgments in evidence-seeking designs is that these designs contain *deontic modals*, that is, words like "has" or "need" (as in the experimental materials quoted just above); and it is uncontroversial that what's at stake should influence what we think someone needs to do in a given context or situation. According to Buckwalter and Schaffer, the influence that stakes have on these kinds of deontic modal claims is what is driving the results, and so the results don't really show that stakes influence knowledge. If these theorists were confined to their armchairs, perhaps we would have only another locus of stalemate. On the armchair model of "explaining away," Buckwalter and Schaffer could have offered

this counter-explanation without any dialectical costs. But since this is a debate between experimental philosophers, it could and did proceed productively. On a further study, Francis and her colleagues were able to replicate the effect, but using prompts that did not include deontic modals: "How many times can Elaine consult her checklist and still not know that she is making the vaccine correctly?" (p. 440). Since they still found a stakes effect on this version of an evidence seeking task, even without any such modals, the Buckwalter and Schaffer hypothesis is badly disconfirmed, defusing it as a means of premise denial in a way that would have been impossible from the armchair.

This active literature on stakes effects in folk epistemology thus highlights three different ways in which X-phi can help us make philosophical progress beyond the bounds of the armchair. First, recall that in Chapter 4, we identified one failure mode for inquiry to be when a falsehood gets calcified into an impregnable dogma. And we can thus see here the potential for X-phi to intervene when armchair philosophy is suffering from a situation of the sort we have quoted Williamson warning about, in which a "few resultant errors easily multiply to send inquiry in completely the wrong direction." We can see anew in our terms here why such consensus errors would be especially damaging to inquiry, as providing a fixed point for debate, but a wrong one—doomed to steering our arguments perpetually the wrong way. If, as seems increasingly likely, epistemologists turn out to have been substantially mistaken about the influence of stakes on knowledge verdicts, then the benefits to inquiry will be correspondingly substantial. Second, X-phi discovered a novel philosophical phenomenon of a stakes effect—not on verdicts, as it seems epistemologists may have mistakenly thought, but rather on the distinct folk epistemological task of deciding when to seek more evidence. This would not have been an easy phenomenon to uncover from the armchair, most particularly because unlike a binary "knows/doesn't know" question, the evidence-seeking experimental design asks for a number, and one would expect individual philosophers to demonstrate (as the experimental participants did) a significant amount of variation in exactly what numbers they might respond with to any given version. It required quantitative methods, using data gathered from an adequately sized sample, in order to detect and confirm this pattern. Third, we have a lovely example of X-phi enabling a direct response to an "explaining away" attempt, which likely would have resulted in a dialectical stalemate under pre-X-phi rules of engagement.

In addition to further testing and perhaps extending the current trends regarding these phenomena, X-phi will be needed to address at least two interesting residual questions: what accounts for this difference between knowledge verdicts on the one hand and evidence-seeking judgments on the other? And why are we seeing what appears likely to be a stark case of demographic variation, namely, philosophers versus non-philosophers, regarding stakes effects on knowledge verdicts? A sizable number of epistemologists have sincerely reported their stakes-sensitive

verdicts on scenarios like DeRose's bank cases, and thus one question that would need to be addressed is how widespread is this response within epistemologists and/or philosophers on the whole? Is there a distinct subpopulation of the profession that displays this effect, or is it something more broadly shared? One cannot read this off the published literature, or through informal conversations, even numerous ones. As X-phi helps us become more deeply informed about the nature of these phenomena, it can then inform our answers to questions about the further philosophical upshots. Epistemologists will want to know: can we find good reason to privilege the verdictive phenomenon over the evidence-seeking one; or vice versa; or if neither, then what does this all mean for the theory of knowledge? Psychology cannot answer that question for us philosophers; just as much so, armchair philosophy cannot answer it either.

5 Controlling for Demand Characteristics in Debates about Personal Identity

One of the central questions in metaphysics involves the nature of *personal identity*. There is longstanding disagreement about what it is that makes us who we are and what kinds of changes can we survive, and we can divide the philosophical landscape roughly into two camps. According to members of the first camp, which traces its roots back to Locke, personal identity depends essentially on *psychological continuity*, or the persistence of certain psychological characteristics (see, e.g., Shoemaker 1963, Perry 1972, Lewis 1976, Parfit 1984, and Unger 1990). According to members of the second camp, which traces its roots back to Butler and Reid, personal identity depends essentially on something else (see, e.g., Williams 1970, van Inwagen 1980, and Swinburne 1984). In "The self and the future," Bernard Williams (1970) famously argued that one reason, perhaps *the* reason, for this divided landscape has to do with how thought experiments about personal identity are framed. Thought experiments framed one way lead us to think that what really matters is the persistence of psychological characteristics, while thought experiments framed another way lead us to think that people can survive the loss of their distinctive psychological characteristics.

To illustrate this, Williams asked readers to think about two different hypothetical cases. In the first case, readers were asked to imagine a scenario in which two people, *person A* and *person B*, will undergo a procedure that transfers all of the psychological characteristics that had been associated with *person A* into the body that had been associated with *person B*, and all of the psychological characteristics that had been associated with *person B* into the body that had been associated with *person A*. The wrinkle is that, before they undergo the procedure, *person A* and *person B* are both told that once the procedure has been performed, one of them will be given a large sum of money and the other will be tortured. Williams

contends that it seems natural to say that after the procedure has been conducted *person A* and *person B* have swapped bodies. This is supposed to be especially obvious when we consider what kinds of special requests *person A* and *person B* might make prior to when the procedure will take place regarding who will be given the money and who will be tortured: each person would, presumably, request that the money be given to the other *body-person* and the torture be administered to the same body-person; that is, each person would want the reward to come to them, and similarly would want to avoid punishment, after the transfers have taken place and they have arrived in their new bodies. And this is supposed to support the idea that personal identity depends essentially on the persistence of psychological characteristics. In the second case, readers were asked to imagine a quite different scenario; one in which they themselves play a starring role. This time, they were asked to imagine that they themselves will undergo a procedure that will change all of their psychological characteristics just moments before the bodies that they currently occupy are tortured. If what we have learned from thinking about the first case is right, and what really matters for personal identity is psychological continuity, then readers should be completely unphased by what's about to happen. After all, with all their customary psychological states temporarily removed, they themselves won't be there to experience the torture. Yet Williams claims, plausibly, that it is natural to think that people will be afraid about what is about to happen to *them*, and this is supposed to support the idea that personal identity does not depend essentially on the persistence of psychological characteristics.

The fact that what we think about personal identity seems to depend on how thought experiments about personal identity are framed is evidence for some metaphysicians that there are significant limits to what we can do using the method of cases. For example, Ted Sider (2001) writes,

> It appears that we are capable of having either of two intuitions about the case, one predicted by the psychological theory, the other by the bodily continuity theory. A natural explanation is that ordinary thought contains two concepts of persisting persons, each responsible for a separate set of intuitions, neither of which is *our* canonical conception to the exclusion of the other . . . I should say that although I claim that use does not favor either the psychological or the bodily continuity theory, I make this claim only tentatively. Perhaps new thought experiments will be devised that tell decisively in favor of one theory or the other. Or perhaps new theoretical distinctions will be made that will make clear that one or the other competing sets of intuitions were confused, or mislabeled. (Recall the effect of Saul Kripke's (1980) distinction between epistemic and metaphysical possibility on intuitions about the necessity of identity.) I doubt these things will occur, but it is impossible to know in advance what future philosophical investigation will reveal. (p. 198)

It is important to note here that Sider's pessimism is framed against the background of what would be needed in order to make philosophical progress toward resolving debates about the nature of personal identity. Where Sider is doubtful, Shaun Nichols and Michael Bruno (2010) are optimistic, actively seeking an empirically grounded way to explain away one or the other of the conflicting verdicts. They write,

> if we can discredit or find additional support for the source of just one of the intuitions, then this might help us in determining whether or not persistence depends on our psychology. More generally, until we know more about what influences these responses, such pessimism seems premature. What factors generate the different intuitions? Why do we get pulled in different directions? What influences which direction we go? Answering these questions may well illuminate which candidate theory of personal identity is more plausible. (p. 297)[23]

They set out to see what progress can be made by looking carefully at *why* people have the intuitions that they do about these cases. To do so, they ran a series of experiments using cases similar to the two famous thought experiments that Williams used in his paper:

> *Transplant Case #1*
> Jim is an accountant living in Chicago. One day, he is severely injured in a tragic car accident. His only chance of survival is participation in an advanced medical experiment called a "Type 2 transplant" procedure. Jim agrees. It is the year 2020 and scientists are able to grow all parts of the human body except for the brain. A stock of bodies is kept cryogenically frozen to be used as spare parts in the event of an emergency. In a "Type 2 transplant" procedure, a team of doctors removes Jim's brain and carefully places it in a stock body. Jim's original body is destroyed in the operation. After the operation, all the right neural connections between the brain and body have been made. The doctors test all physiological responses and determine that the transplant recipient is alive and functioning. They scan the brain of the transplant recipient and note that the memories in it are the same as those that were in the brain before the operation.[24]

> *Transplant Case #2*
> Imagine that some time in the future your brain has developed a lethal infection and will stop functioning within a few hours. In the emergency room, you are

[23] This argument is a good example of what Weinberg (2016c) called "going positive by going negative." Joshua Knobe (2021) has recently signed on to this way of thinking about how experimental philosophy can help us make philosophical progress.

[24] Nichols and Bruno borrowed this case from Blok et al. (2005). It's always fun to see what people in the past thought technology would look like in the future, especially when the future that they imagined is now in the past!

alert and listening as the doctors explain to you that the only thing they can do is the following:

> Render you completely unconscious, and then shave your head so that they can place electrodes on your scalp and shock your infected brain. Unfortunately, this procedure will permanently eliminate your distinctive mental states (including your thoughts, memories, and personality traits.)

You slip into unconsciousness before the doctors can discuss the matter further with you, and they elect to perform the procedure. It works exactly as expected. Several days after the procedure, the doctors perform some follow up brain scans and administer a series of painful shots.

What they found confirmed what Williams conjectured about his two cases. Participants who were asked to think about the first transplant case said that Jim survived the procedure, and those who were asked to think about the second transplant case said that they had as well.

So, what's going on here? Why does it seem natural in some cases to think that personal identity depends on the persistence of psychological characteristics and natural to think in other cases that it does not? Williams offered two possible explanations. First, he suggested that the two cases might have what we now call different *demand characteristics*. In the context of the social sciences, demand characteristics are subtle cues that make participants aware of what the experimenters expect to find or how participants are expected to respond, and the problem is that they can change the outcome of an experiment because participants alter their behavior to conform to expectations. In this context, the first case that Williams discusses in his paper explicitly talks about people *exchanging bodies* and this might naturally lead readers to form the idea that what matters for personal identity are the sorts of characteristics that can, accordingly, move from one body to another, such as, crucially, psychological ones. And the second case that Williams discusses in his paper begins by asking people to imagine themselves in a scenario in which someone tells them that *they* are going to be tortured and this might naturally lead readers to form the idea that whatever else happens before tomorrow comes, *they* will be there when it does. And second, Williams suggested that perhaps *perspective* matters. In the first case, readers are asked to think about a procedure being performed on *other people*; in the second case, readers are asked to think about something that is happening to *them*.

While Williams was left to speculate about possible explanations for this framing effect, and Sider was left unable to make armchair progress at this specific locus of dispute, Nichols and Bruno were able to test these potential explanations, and ultimately to experimentally disconfirm one of them. They repeated the experiments just rehearsed using both second-personal and third-personal versions of the two

cases, and found, contrary to Williams' second proposed explanation, that perspective does not seem to matter. Participants who were presented with the first transplant case continued to believe that what really matters is the persistence of psychological characteristics regardless of whether they were asked to think about a second-personal version of the case or a third-personal version of the case. And participants who were presented with the second transplant case continued to believe that people can survive the loss of their distinctive psychological characteristics regardless of whether they were asked to think about a second-personal version of the case or a third-personal version of the case. This leaves the first potential explanation, that is, that demand characteristics cases cause people to think that personal identity depends on the persistence of psychological characteristics, while other cases cause people to think that it does not. While additional experimental work would be needed to establish whether that explanation is correct, what matters, for our purposes here, is that researchers could significantly disconfirm one of the potential explanations. Nichols and Bruno's work succeeds in advancing the dialectic by raising the cost for anyone trying to stick with a perspective-based account of the Williams effect; and they do this using tools that can only come along with X-phi.

Remember how this subsection began, namely, with the philosophical landscape divided into two camps: those metaphysicians who believe, like Locke, that personal identity depends essentially on psychological continuity and those who believe, like Butler and Reid, that it depends on something else. Williams speculated that the divided philosophical landscape was due to the fact that the cases that philosophers typically use to support their preferred theories of personal identity frame things in such a way that their own preferred theories come out to be true.[25] And experimental work confirms this hypothesis: thought experiments framed one way seem to lead us to think that what really matters is the persistence of psychological characteristics, while thought experiments framed another way lead us to think that people can survive the loss of their distinctive psychological characteristics. But, more than this, experimental work also helps us begin the process of weeding through possible explanations for this framing effect. Progress in this second endeavor, that is, in the attempt to determine why framing matters in the kinds of cases typically used in debates about personal identity, comes precisely by bringing the relevant philosophical phenomenon into sharper focus, that is, by separating signal from noise, as Bogen and Woodward suggest.

[25] Recent experimental work shows that a similar explanation can be given for a similarly divided landscape in the epistemology of disagreement (see Alexander et al. 2018). In particular, this work demonstrates that two different framing effects influence how people think about many of the cases that have been used in the epistemology of disagreement, and that what people think about these cases tracks whether the cases have been used to support *steadfast* or *conciliatory* theories of peer disagreement.

6 Dislodging Mistaken Empirical Assumptions in Debates about Reference

In our section on pragmatic encroachment, we saw an example of X-phi perhaps disturbing a mistaken consensus view about a piece of candidate phenomena. Our final example of the progressive potential for X-phi will exemplify a similar sort of progress, by rejecting a mistaken view of what the space of acceptable theories must look like.

Our case study comes from the philosophy of language and ongoing debates about the reference of proper names. Two broad families of theories have dominated: *descriptivist theories*, according to which names refer to whatever best satisfies the descriptions that competent users of those names associate with those names (Frege 1893, Russell 1919, Searle 1958, Lewis 1972), and *causal-historical theories*, according to which names refer to whatever was picked out when those names were first introduced (Geach 1969, Donnellan 1970, Putnam 1975, Kripke 1980).

Much of the debate about theories of reference has involved the method of cases—in both armchair and experimental varieties. As Shaun Nichols, Ángel Pinillos, and Ron Mallon (2016) observe,

> Much of the work on theories of reference (both armchair and experimental) proceeds on the assumption that theories of reference must accommodate competent speakers' judgments concerning various classes of terms—proper names, natural kind terms, demonstratives—in various actual and possible circumstances. For example, if a theory of reference entails that in a given scenario, *n* refers to *x*, then the theory is (defeasibly) supported if competent speakers judge that in the scenario, *n* does refer to *x* (or something which implies that judgment); the theory is (defeasibly) undermined if competent speakers make the opposite judgment (or something which implies that judgment). Traditionally, theories (including the authors of classic works on reference in the philosophy of language) have primarily relied on their own judgments, presumably on the assumption that they themselves are representative competent speakers. More recently, experimental philosophers have used experimental methods to measure intuitions about reference. (p. 147)

In Chapter 1, we discussed what is perhaps the most famous example of how the method of cases has been used in this debate, namely, Kripke's (1980) famous Gödel Case, as well as some influential experimental work on these kinds of cases (Machery et al. 2004, Mallon et al. 2009).[26] That initial work was largely aimed

[26] This work has generated a considerable amount of attention, prompting a number of subsequent empirical studies. See, e.g., Ludwig (2007), Deutsch (2009), Marti (2009), Machery et al. (2009), Lam (2010), Devitt (2011), Sytsma and Livengood (2011), Ichikawa et al. (2012), Machery et al. (2015).

at demonstrating cultural differences in such cases, hypothesizing that Western speakers had broadly casual-historical verdicts but Eastern speakers more descriptivist ones. But here we want to focus on more recent work by Shaun Nichols, Ángel Pinillos, and Ron Mallon (2016) that suggests that instead of two different theories of reference each working primarily for different groups of people, perhaps we all share the same access to both kinds of reference for natural kind terms (and, they suggest, perhaps names as well). Reference might simply be *ambiguous*—that is, that it is sometimes natural for anyone to think about such terms descriptively and sometimes natural to think about them causal-historically, and that which way of thinking about them is the right way depends not so much on where they grew up, but rather on the conversational setting they find themselves in. One part of the motivation for this view comes from the way that philosophers engaged in debates about theories of reference have used the method of cases:

> Broadly speaking, our strategy for developing different manipulations relied on the kinds of philosophical arguments used to defend descriptivist and causal-historical theories of reference. The idea is that in defending one or the other theory of reference, philosophers deploy presentations and examples that guide our intuitions to favor the author's preferred theory. (p. 151)

Another part of the motivation for this view comes from Philip Kitcher's (1978, 1993) influential account of how theoretical terms are used in the history of science, where Kitcher argues that the language used to present scientific theories sometimes prompts us to think about theoretical terms descriptively and sometimes prompts us to think about theoretical terms causal-historically.

To demonstrate that reference is ambiguous, Nichols and his colleagues ran a series of experiments drawn from this history of science, and in particular from medieval bestiaries, which were collections of short descriptions about various animals, real and imaginary. Here we will focus on one of them. In this experiment, participants were asked to read the following passage:

> In the Middle Ages, animal researchers described a distinctive kind of mammal. They called it *catoblepas*. The catoblepas was said to be like a bull but with a head so heavy that the animal has to keep its head down at all times. It was also thought that the catoblepas had scales on its back. In addition, the researchers said that looking into the animal's eyes causes immediate death. Of course there is nothing that meets this description, but researchers know that it was based on reports of encounters with wildebeest. Many scientists in the middle ages, when faced with wildebeest, would often call them catoblepas.

After reading this passage, participants were asked the extent to which they agreed with the claim that catoblepas exist and the claim that catoblepas are wildebeests.

What they found is fascinating. Participants were willing to say that catoblepas are wildebeests but unwilling to say that they exist. This suggests that what participants think that they are talking about when they use the term 'catoblepas' changes depending on the kind of question that they are being asked to answer. When participants are asked whether catoblepas are wildebeests, they are confronted with a question that presupposes that catoblepas exist and they accommodate that presupposition by adopting a causal-historical interpretation of the term 'catoblepas'—after all, since they have just read that nothing fits the description associated with the term, whatever we are being asked to talk about, it has to be whatever it was that people were talking about when they first used that term, namely, wildebeests. So, in that kind of context it is natural for them to think that when we talk about catoblepas we are just talking about wildebeests. By contrast, when participants are asked whether or not catoblepas exist, it's an open question whether or not they do, and so no accommodation is needed, at least not the kind of accommodation that seems to push participants to think about catoblepas causal-historically. In this kind of context, participants seem to approach the question descriptively, and since the passage that they've just read tells them that nothing fits the description associated with the term 'catoblepas', it is natural for them to say that catoblepas don't exist.[27]

While this is only one study, and thus much too early to declare its results as established, it can still serve to exemplify a kind methodological benefit for how we think about the longstanding debate about theories of reference, should the work be replicated and extended.[28] As Nichols and his colleagues note,

> This suggestion runs against just about everything in the literature. In the philosophy of language, work on the theory of reference has operated under an assumption that only one theory of reference will apply to a class of terms. This assumption has been a critical constraint on theory building in philosophy of language . . . Our ambiguity theory thus abandons a dominant line of thought across a wide swath of philosophy. (pp. 160–161)

What we want to suggest here is that it also points to another way in which experimental work can help us make philosophical progress, namely, by revealing that a longstanding philosophical disagreement can rest on an empirical mistake. The

[27] We are adopting the language of "accommodation" here, but Nichols and his colleagues are quick to point out that the ambiguity they've uncovered cannot be explained entirely in terms of accommodation. In fact, they do not think that there is any comprehensive story to be told that determines when it's right to think about terms descriptively or causal-historically. They take consolation in the fact that this makes this kind of ambiguity similar to other kinds of linguistic ambiguity, where ambiguous terms refer to whatever speakers choose to talk about given their conversational goals. Nonetheless, we do not take any view here on this further matter.

[28] As one would expect, there have been both critical and supportive follow-up studies by other researchers as well. See in particular Martí (2020), Devitt and Porter (2021), Haukioja et al. (2023).

philosophy of language has been mired in debates about theories of reference that have rested on the assumption that reference is fixed either one way or the other. Let's suppose that the authors are right, in their claim that inquiry in the theory of reference has widely been presupposed to be theoretically univocal for any given class of terms. If that is so, then their experimental work would have helped reveal that this assumption is mistaken, and dislodged it from where it may have been blocking progress in this domain; if how reference is fixed generally depends on the context in which people find themselves, then it is not univocal at all. In some conversational contexts, it is natural to think about names descriptively, and in other conversational contexts, it is natural to think about them causal-historically. Since how we should think about names would depend on the conversational setting we find ourselves in, and since we find ourselves at different times in different conversational contexts, there would be no one way that we should think about reference. The debate, it seems, would perhaps be no debate at all.

7 Standing Up for Philosophical Progress

We have been suggesting here that one significant step toward greater progress in philosophy is for us to bring philosophical phenomena into sharper focus, characterizing the noteworthy patterns of philosophical matters at a remove from the data of specific philosophical cases and other sorts of "observables" that we perhaps get mustered too quickly into premises for our arguments in a way that tends to leave our opponents in a position of costless denial. To underscore X-phi's ineliminable role in our achieving this kind of progress, we will close by critically considering an attempt to lay claim to a similar form of progress from the armchair. Is it truly the case that more of this kind of progress *couldn't* nonetheless be made from the armchair? Perhaps with just a bit more time and comfortably seated effort? After all, *some* of the things that we have been suggesting that experimental methods can help us do are not completely unknown to philosophers engaged in armchair philosophical practice. Here is an argument from David Papineau (2009) putting forward a closely similar role for addressing conflicting case verdicts in armchair analytic methodology:

> Go back to the idea . . . that philosophy is characteristically concerned with theoretical tangles. We find our thinking pulled in opposite directions and cannot see how to resolve the tension. Often part of our predicament is that we don't know what assumptions are directing our thinking. We end up with conflicting judgments, but are unclear about what led us there. In such cases thought-experiments can bring the implicit principles behind our conflicting judgments to the surface. They make it clear what intuitive general assumptions are governing our thinking and so allow us to subject these assumptions to explicit examination. (pp. 22–23)

So, the story might go, even if experimental tools and methods could provide us with tools that *might* help us make some sort of philosophical progress, we nonetheless already possess a rubric for how to pursue progress without *having* to buy what the experimentalist is selling.[29] The idea is that armchair methods can detect reasonable systematic variation across cases, and extract the implicit principles in them. And then, once subjected to the kind of "explicit examination" that Papineau has in mind, some will turn out not to be worth continued endorsement and the case verdicts that these positions had enabled us to make will fall by the wayside, as Ichikawa unpacks in terms of "explaining away." According to this story, this kind of explicit philosophical reflection can provide better insight into our previous philosophical commitments and why the philosophical positions we now endorse merit that endorsement. What's more, uncovering hidden assumptions is exactly the sort of thing we might well expect to be able to do from our armchairs, in part because the conceptual space of available philosophical positions can be determined by thinking about what might plausibly make sense for us to assume. After all, the set of available philosophical positions tend to be ones that have at least some *prima facie* plausibility in their own right.

This is an attractive story for armchair philosophy, but as we have argued throughout the course of this book, experimental philosophy teaches us that many "philosophical tangles" will simply not derive from anything so respectable, so rationally discernable, so tractably *principle-like* as the "assumptions" that Papineau has in mind. This is one of the most important lessons that philosophers can learn from paying close attention to the social and cognitive sciences: a nontrivial number of philosophical tangles simply cannot be spotted from the armchair, let alone successfully untangled using the tools and resources that armchair philosophers have available to them.[30] Most obviously, when the variation is one between philosophers and non-philosophers, this will be very hard for the philosophers to spot from unsystematic observation; this may be the case with stakes effects on knowledge verdicts. And as we discussed in Chapter 5, subtle context and framing effects are undetectable to introspection or unsystematic observation, and demographic differences are obscured both by selection effects, where perhaps only those philosophers with the "right" verdicts will advance in specific domains of philosophical inquiry, and by the plain fact of professional philosophy's lack of demographic diversity. This is one of the reasons why, we argued, the error profile that philosophical folk wisdom gives us about the method of cases just doesn't provide us with nearly complete enough an accounting of what kinds of things really will or won't cause problems when philosophers use the method of cases.

[29] Appiah (2008) makes a similar suggestion, and interestingly so does the psychologist Tania Lombrozo (2013), who is otherwise a proponent of experimental philosophy. For additional discussion of Appiah's version of this response, see Weinberg and Wang (2010).

[30] A similar argument is made by Schickore (2019), in a broader discussion of controls and confounds, and Craver and Dan-Cohen (2024), in a broader discussion of artifacts and experimental design.

And we have seen that this is especially true for the kinds of philosophical tangles that we have examined in this chapter, where untangling these seemingly Gordian philosophical knots required looking for things other than narrative misunderstandings, modal confusions, or philosophical bias—the kinds that are included in armchair philosophy's error profile for the method of cases, and the kinds of things that we can look for without having to use experimental tools and methods. As we've seen, straightening out the kinds of philosophical tangles that we have been looking at in this chapter requires us to be able to understand things like complex psychological processes, how to distinguish artifacts from real effects, and the non-rational role that demand characteristics play in shaping how people think about philosophical cases—the kinds of things that are *not* included in armchair philosophy's error profile for the method of cases, and the kinds of things that we have to stand up to look for using experimental tools and methods. What this means is that, while Papineau's armchair-friendly philosophical tangles may well deserve to be unwound and unraveled, with their various entwining threads perhaps being worthy of philosophical examination in their own right, there are important philosophical tangles that cannot be untangled in that way—ones that can only be untangled using experimental tools and methods.

This brings us to one last methodological benefit that we wish to present here of philosophizing with our feet firmly on the ground. We emphasized above that a flaw, in fact, one of the fatal flaws, in our armchair practices is how we allow premises to be denied on the basis of merely plausible hypotheses; and that armchair resources are often inadequate to push beyond the plausible to the actual. Experimental resources don't just let us ultimately decide between rival plausible hypotheses, though; on occasion, they can also lead us to endorse hypotheses that were initially *implausible*, as many of the quirkier sorts of unconscious mechanisms may strike many philosophers. This is nicely illustrated by an ambitious research project by Eugen Fischer and his collaborators, in which they are trying to pin a psychologized version of Austinian quietism on the theory of perception. Their hypothesis, in a nutshell, is that many conundrums about perception are artifacts of our being prey to *unconscious default inferences* about words like "see," which philosophers are unable to overcome; for example, that from a sentence of the form "S sees A" one can infer that there is some entity A that is being seen by S, and thus in cases of hallucinations, there must be some *thing* that is nonetheless seen (like MacBeth's dagger, or perhaps, "dagger"). Now, we expect that most philosophers will find it not particularly plausible that we could continue to be suckered by such a default rule of inference that we know to be a mere default, especially when we are in the fully reflective and attentive crucible of professional philosophical activity. But, over the course of a dozen papers, drawing on a wide range of empirical methodologies, including eye-tracking experiments, and amassing a convincing amount of evidence, Fischer and his colleagues have demonstrated that this really is how our minds actually work. (See Fischer 2023 for a

recent summary.) While we think the jury is still out on their attempt to dissolve many puzzles in the philosophy of perception, their project shows us how it may be possible for experimental philosophy to go well beyond the armchair in devising, and confirming, hypotheses about our philosophical cognition that fall well outside of philosophical common sense.

We think that there's a more general lesson here to be learned from all of this, one that underscores why it is so important for us to stand up more for philosophy. What is a methodological virtue when we are trying to decide hidden principles threatens to be a troublesome vice when we must seek more oddball distorting factors in our cognition. Philosophers have a professionally trained bias toward claims that make sense, that are at least plausible candidates for enthymemes, and are worthy of some extended examination and consideration. Those are the sorts of things one can do some philosophy on, perhaps along the lines articulated by Ichikawa. But psychological heuristics and biases and the quirky influences of factors that lie beyond any plausible rationalizing story—these are not generally going to occur to philosophers on a Papineauvian project. Many of them, like order effects, don't even vaguely resemble premises that could be subjected to rational scrutiny, and will not be likely to be hypothesized in Papineau's proposed procedure. Our education and experience perhaps cultivate in us a keener sensitivity to one sort of unconscious cognition, but at the cost of a diminished capacity to notice or theorize in terms of other sorts. Philosophers, thus, suffer a bit of *déformation professionnelle*. It turns out to be very hard to turn ourselves into finely crafted instruments that can perform some specific set of tasks with great discrimination and puissance without rendering ourselves at the same time rather clunkier when it comes to still other tasks that are no less subtle and important but for which we are now dangerous ill-suited. As we will argue in the next chapter, the key to responding to this will be *professional reformation*.

7
Putting Philosophy Back On Its Feet

So analytic philosophy has fallen into a state of methodological irrationality. We are apportioning our methodological resources in a way that leaves us unnecessarily and expensively vulnerable to a wide array of possible errors. As a matter of methodological cost-benefit analysis, we would do better to make a substantial transfer of time and energy into error-management activities, and the requisite sorts of activities do not appear to be ones that can be adequately pursued from within the artificially ascetic strictures of the armchair. Thinking in terms of this cost-benefit argument, in the previous chapter we argued that leaving the armchair offers a potential boon to philosophical progress, adding a range of significant philosophical benefits to that side of the ledger, including but going well beyond establishing particular verdicts as robust. In this chapter, we take up the matter of costs: just what sorts of expenditures of resources do we have in mind? We will argue that while the costs of this transfer would be real, and would take serious and intentional efforts on the part of our community to implement, they nevertheless are nowhere near as ruinous as we think many armchair practitioners may fear. The philosophical expenditures turn out to be quite affordable, easily outweighed by the expected philosophical gains.

We will consider the question of methodological costs at several levels of abstraction. We will begin by articulating a role for experimental methods in philosophical practice that avoids, on the one hand, being so exorbitant that it may look like the present, actually lived methodological costs would swamp the future, merely potential benefits, and on the other hand, standing at such a specialized remove from mainstream philosophical practices that the addition of experimental methods could not provide us with the needed help with error avoidance and mitigation. After we have done that, we will attempt to sketch a series of rather more concrete policy proposals for how we can modify philosophical practice in order to bring it into compliance with the requirements of methodological rationality; these proposals will include proposed revisions to graduate education and philosophical publication practices, and an expansion of our inferential norms. Finally, we will sketch a picture of what the method of cases could look like once we've stood up from our armchairs, where this doesn't involve simply turning philosophy wholesale over to psychologists or suggesting that we philosophers must all become ones ourselves. As we hope to have made clear throughout, we are

Standing Up for Philosophy. Jonathan M. Weinberg and Joshua Alexander, Oxford University Press.
 DOI: 10.1093/9780191947704.003.0007

reformers, not revolutionaries,[1] and our aim in this last chapter is to help philosophers see how easily they could themselves inhabit and implement the needed modifications to philosophical practice. While we, of course, would love for the recommendations that follow to be accepted and implemented, our more realistic hope is that the proposals that we make here will spark meaningful discussions within the philosophical community about the ways in which we can improve our methods. One of our primary goals throughout the book has been to demonstrate that philosophical *practices* can, indeed *must*, be a target of metaphilosophical inquiry in its own right, and so our aim here is to begin what we hope will be a constructive way of engaging in this kind of inquiry.

1 Experimental Philosophy and Philosophical Practice

Our friend Ron Mallon, who was a vegetarian for many years, had a line about why many people simply couldn't imagine themselves as vegetarians: such a person would, he thought, imagine a steak dinner, with a potato and maybe a side of broccoli, and then simply subtract the steak from this mental image, leaving a clearly inadequate dinner in their stomach's mind's eye. And they'd ask themselves, who could live like that? While we take no particular stance here on the ethics of eating animals, the broader moral of this story generalizes: if you want someone to willingly give something up, you'd better help them understand what you want them to replace it with, and it better be something both appetizing and nourishing. And so we had better have something better to offer philosophers at this point in our story than the philosophical equivalent of a baked potato and a side of broccoli, lest the entire story that we have been trying to sell philosophers on prove too hard to swallow. And so we think that it is perfectly appropriate at this point for analytic philosophers to ask, if we are supposed to leave our armchairs behind, then how are we supposed to furnish our workplaces?

The Universal Experimentation Model

The most obvious answer, or at least the one that analytic philosophers are often worried is the most obvious answer, is to replace our armchairs with *lab benches*, where analytic philosophers are supposed to replace traditional philosophical activity wholesale with systematic experimental investigation. On this way of thinking about the future of philosophy, *everywhere* that philosophers might once have appealed to for verdicts about cases is now supposed to be a place where

[1] We are lovers, not fighters.

philosophers, or perhaps psychologists, do experimental work to establish the robustness of those verdicts across demographic or situational variations. Let's call this the *universal experimentation model* for the role that experimental methods would now play in philosophical practice. While this is perhaps the most conceptually uncomplicated successor to the armchair philosophical practices, and while adopting this model would surely bring a significant number of error-management resources into play in philosophical inquiry, we would reject this way of thinking about the future of philosophical practice. Quite simply: it would cost us too much.

To see why, let's consider what a parallel universalization of *formal* techniques would look like in philosophical practice. After all, philosophers throughout history have sometimes committed formal fallacies in their arguments. Surely, this happens less frequently in contemporary analytic philosophy, but it still happens with nontrivially greater than zero frequency that philosophers will trip over a scope ambiguity or an unintentional inversion of quantifier order. As we noted in Chapter 4, Williamson suggests that a major way that philosophers can "do better" would be for us all to pay closer attention to matters of language and logic. Let's imagine what things would look like if a band of metaphilosophical zealots tried to take his suggestion much, much too far: suppose we were to alter our practices to require both the rigorous translation of *all* philosophical arguments into an appropriate formal language and the formal derivation of conclusions from premises, perhaps together with additional assumptions. Such a hyper-formalist revolution would surely reduce the number of fallacious philosophical arguments, maybe even nearly eliminate them altogether. There would, in other words, be some tangible benefits from the perspective of error management. Nevertheless, the costs associated with implementing these kinds of methodological norms would swamp the value gained by decreasing, or even eliminating, these kinds of fallacies. And the costs here are not just practical costs associated with graduate training or journal refereeing; they are alethic as well. Even if we assume that errors associated with performing formal derivations could be caught prior to publication, which is not altogether implausible to suppose, nonetheless there is an increased risk of introducing philosophical errors during the translation to or from the formal calculus, something all too familiar to anyone who has spent time thinking about, say, the various and varied ways that logicians have proposed capturing the meaning of conditionals when moving from natural to formal languages. What's more, this kind of methodological norm would deprive us of the contributions of philosophers who are otherwise insightful and skilled, but who lack these specific technical chops.

All in all, we would be curtailing one vector of errors by introducing still worse ones, and for a benefit that would wildly fail to offset the cost. Even the strongest proponents of formal methods in philosophy would, we expect, agree that such a "universal formalization model" would not be wise for philosophers to adopt as a guide to how formal techniques should be used in philosophical practice. And

returning to the matter at hand, it seems to us obvious that the universal experimentation model would fare at least as poorly as its formal counterpart, and for all the same reasons *mutatis mutandis*. Such a model falters upon the uneven distribution of the relevant aptitudes in the profession; the increased risk of error at the stage of operationalization and design of materials when trying to test those claims that are not especially amenable to experimental treatment; substantially higher practical costs, since it is generally much more expensive to run a good study with adequate power than it is to work out a formal proof; and so on. For these reasons, we intend in this chapter to identify and promulgate a less totalizing role for experimental philosophy.

For what it's worth, while many traditionally minded philosophers have taken experimental philosophers to be advocating for this kind of radical view, in general this has just not been so. For example, in our 2006 paper on order effects, we ourselves, with our colleague Stacey Swain, only proposed "that philosophers who wish to continue relying on intuitions as evidence *begin* empirically investigating intuitions about their favorite thought-experiments to determine whether, and which, intuitions may be taken as evidence" (pp. 153–154; emphasis added). Edouard Machery and Stephen Stich (2013) do seem at first to endorse something like the universal experimentation model in their paper on methodology and the philosophy of language, when they argue for "philosophers of language to follow the lead of generative syntacticians and to replace the appeal to their own and their colleagues' intuitions with systematic empirical studies of ordinary speakers' intuitions" (p. 499). But, while that "replace" certainly looks totalizing, a few pages later they soften their position substantially and end up only objecting to the "*exclusive* reliance on their own and their colleagues' intuitions" (p. 504; emphasis added).

More irenic theses are much more common. For example, Joshua Knobe and Shaun Nichols (2008b) clearly state that the goal of their "manifesto" is to "make clear the nature of experimental philosophy, as well as its *continuity with traditional philosophy*" (p. 4; emphasis added). They go on to stress that

> No one is suggesting that we boot out all of the moral philosophers and replace them with experimentalists, nor is anyone suggesting that we do away with any of the methods that have traditionally been used for figuring out whether people's intuitions truly are right or wrong. What we are proposing is just to add another tool to the philosopher's toolbox. That is, we are proposing another method (on top of all of the ones that already exist) for pursuing certain philosophical inquiries. (p. 10)[2]

[2] Let us foreshadow the next section by noting one key divergence between our proposals and the tenor of Knobe and Nichols' proposal here: rather than thinking of X-phi as just one more tool *in parallel* with traditional, armchair tools, we are urging the *integration* of X-phi practices more broadly into our practices with the method of cases in particular.

The continuity of experimental and traditional philosophy has also been championed by those more traditionally minded philosophers who are nonetheless comfortable with psychological methods and experimental philosophy itself. Thus does Jennifer Nagel (2012) write about the relationship between experimental philosophy and analytic philosophy in her prominent defense of the method of cases:

> I agree with the experimentalists that the question of the epistemic status of epistemic intuitions is an excellent question, and that we can take some steps towards answering it by taking a close look at the empirical facts about intuition. (p. 496)

> Fortunately, in epistemology as in physics, intuition is not the only tool at our disposal: considerations of theoretical unification can also supply some guidance. We suspect sensory illusion where the deliverances of the senses appear to conflict with one another, as in the Müller-Lyer illusion; we may have similar suspicions where there is apparent conflict among our epistemic intuitions, for example, conflict of the sort found in the cases motivating contextualism. It is not transparent that these apparently conflicting intuitions are illusions; theorists of various inclinations have developed innovative and sometimes strange theories of knowledge and knowledge ascription on which the apparent conflict is no more than apparent. To judge whether these theories are true or false, we can draw on a great array of considerations from logic, linguistics, psychology and philosophy; we can also devise new cases to offer positive support to our theories or to serve as counterexamples. In epistemology, as in empirical science, it is not always a trivial matter to determine whether we are subject to an illusion, or whether the phenomenon we are investigating is stranger than we had thought. But the fact that a form of inquiry is difficult does not entail that there is anything fundamentally wrong with its methods. (pp. 521–522)

Where we have disagreed with Nagel, here we only disagree about the implicature that she perhaps expresses in the final sentence, that there is nothing fundamentally wrong with the armchair method of cases. As we discussed in Chapter 3, Nagel is primarily concerned with questions of baseline accuracy of verdicts about philosophical cases, and because her focus is on reliability, she is comfortable taking the position that current philosophical practices are more or less in good order as they stand. In her view, although these practices can be usefully supplemented by experimental work, such work would be methodologically supererogatory. The method of cases, she would say, is already "reliable enough." However, as we argued, once we recognize that baseline accuracy is insufficient for methodological trustworthiness, it becomes clear that the method of cases cannot be defended in those terms. There is too much threat of error from too many possible sources, and we are still at such an early stage in these investigations that mostly what we are learning is how much we have left to learn, with new sources of error popping up all the time. This means

that the method of cases *requires* the sort of work that we are advocating for here, and on a fairly substantial scale. Nonetheless, the work is meant to supplement and complement more traditional philosophical practice, not replace it.

All of this means that, while we do not blame anyone for thinking that a universal experimentation model is what folks like us have had in mind, we think that a less radical picture is in fact more consonant with the general tenor of the arguments that even the most die-hard experimental philosophers have had in mind pretty much all along. Of course, this leaves us with the question of what this less radical picture ought to be, and in trying to sketch it, we have to make sure not to err too far in the other direction. If it is truly methodologically irrational for analytic philosophy to under-invest in experimental practices, then we can't arrive at a position where philosophers using the method of cases can almost always just choose to take X-phi or leave it.

The Parallel Subspecialty Model

For exactly that reason, we thus will need to avoid another model of experimental philosophy as an optional addition to the philosophical inquiry, of a sort that has been defended recently by Joshua Knobe (2016) in his influential paper "Experimental philosophy is cognitive science." We will call this the *parallel subspecialty model*. On this model, experimental work on topics like knowledge or moral permissibility would be radically discontinuous with more traditional philosophical work that might march under the banners of epistemology or normative ethics. Instead, as the title of Knobe's article suggests, this model treats experimental work as continuous with, and perhaps totally indistinguishable from, the kinds of psychological or neurophysiological work being done in the cognitive sciences. Rather than studying philosophical cognition with an eye toward better understanding and improving philosophical methodology, and perhaps, as we argued in the previous chapter, with an eye toward better understanding the nature, and possible resolution, of longstanding philosophical debates, experimental philosophers would simply be in the business of studying philosophical cognition for its own sake. As Knobe (2007) writes in an earlier paper about the philosophical significance of experimental philosophy,

> For a proper understanding of the aims of experimental philosophy, we need to adopt a broader historical perspective. It is true that some 20th-century philosophers believed that the main aim of philosophy was to determine the extensions of certain concepts, but this is a relatively recent development. For the vast majority of its history, the discipline of philosophy was assumed to have a far broader purview. In particular, philosophical inquiry was assumed to be concerned in a central way with questions about how the mind works—whether the mind could

> be divided into separate parts, how these parts might interact, whether certain kinds of knowledge were learned or innate, and so on. These questions were then assumed to have important implications for issues in moral and political philosophy. We can refer to this conception of philosophy as the *traditional conception*. It is the conception that was dominant throughout most of the history of philosophy. (pp. 119–120)

A short bit later, he contends that

> With this historical context in place, we can provide a clearer statement of the aims of experimental philosophy. The aim of most work in experimental philosophy is not to answer the new sorts of questions that rose to prominence in the 20th century. Rather, the aim is to address the *traditional* questions of philosophy—the sorts of questions one finds in the work of Plato, Aristotle, Spinoza, Hume, Nietzsche, and so many others [i.e., questions about how the mind works]. (p. 120)

So, on this way of thinking about the relationship between experimental philosophy and analytic philosophy, experimental philosophy is a move away from what Knobe thinks is a mistaken turn toward conceptual analysis and away from traditional questions about how human minds work—the kinds of questions that he says were interesting to philosophers throughout the history of philosophy. And so experimental philosophy takes back up these questions in a manner characteristic of cognitive scientific research. Knobe thinks that this work often leads experimental philosophers toward more general and unifying sorts of psychological explanations, and away from questions about philosophical concepts.[3]

Knobe makes his most forceful case for this way of thinking about experimental philosophy in his (2016) contribution to the Sytsma and Buckwalter *Companion*

[3] It is interesting to note that the trajectory of Knobe's thinking about his own research exemplifies this pattern. In his early papers on what he (with due modesty) calls the *side-effect effect*, or what everyone else (with due respect) calls the *Knobe effect*, he took himself to be discovering features of "the folk concept of intentionality" (the title of his (1997) paper coauthored with Bertram Malle). But over time, as the studies piled up, it became clear that this influence of the positive or negative valence of side effects manifested more generally across human cognition, including how we think about advocacy, causation, choice, decision, desire, knowledge, and preference (see, for example, Pettit and Knobe 2009, Hitchcock and Knobe 2009, Roxborough and Cumby 2009, Beebe and Buckwalter 2010, and Knobe 2010). The Knobe effect has nothing specifically to do with intentionality, after all. It is worth pointing out that while Knobe might be done with the method of cases, the method of cases might not be done with Knobe. That is, on the picture that Knobe sketches, once we have identified that the Knobe effect is a general effect, then we likely will want to *factor it out* of the case verdict data. From a metaphilosophical point of view, it is not just an effect, it is a *bias* (see Nadelhoffer 2004, 2006, and for discussion, Alexander 2012). This brings us back to one of the central points that we made in chapter two, namely, that many of the results that Knobe scores as instances of stability are, from a methodological point of view, noise. In fact, the apparent universality of the Knobe effect is precisely what makes it a source of worrisome contamination for the method of cases.

to Experimental Philosophy. In defending his proposed placement of experimental philosophy within the demesne of cognitive science, here is how he thinks that experimental philosophers should think about contemporary epistemology:

> researchers within the conceptual analysis tradition have spent decades studying intuitions about knowledge. We can now ask what sorts of results this work has delivered. There can be little doubt that it has taught us many interesting things about people's intuitions. That is, it has revealed a number of important and very real effects that are amply worthy of further study. Yet, at the same time, there is a widespread feeling that work on this topic has not converged on anything even remotely resembling a "theory of the concept of knowledge." One natural response to this outcome would be to conclude that this simply isn't the sort of area in which proper theory is possible. For each of the surprising effects that researchers have uncovered, we should of course be seeking deeper theoretical understanding, but there is no reason to demand at the outset that this understanding must come from a theory that has anything to do with knowledge in particular. It might well come from a theory at some other level.
>
> Be that as it may, it seems that contemporary work in experimental philosophy has not been in the business of constructing theories about the use of individual concepts. Thus, if this work is to display theoretical virtues, it cannot manifest those virtues in precisely the manner familiar from the aspirations of conceptual analysis. Whatever virtues it might embody must be understood in a somewhat different way. (pp. 45–46)[4]

A bit further, he underscores the separation between experimental philosophy and forms of inquiry that are characteristic of the analytic tradition:[5]

> Most research in experimental philosophy is so radically different from traditional conceptual analysis that it would be a mistake to think of it as doing anything like what conceptual analysis originally aimed to do. (p. 48)

We are not sure about that "most," but we would certainly agree that *much* experimental philosophy should indeed be understood this way. It's also clear to us that

[4] While Knobe focuses his attention on the relationship between experimental philosophy and *conceptual analysis*, little about his discussion turns on that specific conception of philosophical methodology. What he says here applies to how we have been talking about traditional analytic philosophy throughout the chapter and book.

[5] It can seem confusing to tell whether Knobe is pro- or anti-philosophical tradition. Part of the reason is that he has two traditions in play in his discussion: he is locating his preferred way of thinking about experimental philosophy as continuous with an older philosophical tradition, one that he associates with a number of significant figures in the history of philosophy, while taking the analytic tradition to be divergent from all that. He uses "conceptual analysis" as a name for what is going on in the analytic tradition, although we think it would apply just as well to philosophers who use the method of cases but who do not think of themselves as analyzing concepts per se (e.g., Sosa).

much of it is, nonetheless, very much engaged in being relevant to analytic philosophical inquiry, and to the method of cases in particular; we reviewed several recent examples of this kind of work in the previous chapter.

Having said that, our concern here is not really with the descriptive adequacy of Knobe's view that experimental philosophy is cognitive science, which is truly illuminating about a wide stretch of recent work. Our worry is with the model he sketches of experimental philosophy's relation to analytic philosophy on the whole. If experimental philosophy is *just* cognitive science, puttering away happily in a philosophically remote corner of the naturalistic metaphilosophy universe, then whatever light it may generate to shed on its own theoretical questions will be too far distant to help dispel the armchair's methodological shadows. To do the work we need it to do, we cannot allow experimental philosophy to be *merely* yet another sub-subfield of its own.

Yet there's a very real and pressing question about whether this fate can ultimately be avoided. The prominent philosopher of neuroscience John Bickle (2019) has issued a grave warning to experimental philosophers, like the Ancient Mariner to the Wedding Guest, cautioning against the fate that befell him and his crew of "neurophilosophers." The neurophilosophy movement launched ambitiously in the mid-to-late twentieth century with the promise of revolutionizing philosophy on the whole, with all traditional questions of metaphysics or epistemology to be transformed—and then solved!—by the widespread adoption of neuroscientific concepts and explanations. Bickle succinctly evokes the heady vibes of those early days: "Nascent neurophilosophy: come to learn some cool new science; stay to overthrow perennial philosophy" (4). And yet over the course of a generation, those revolutionary fires were banked, and

> neurophilosophy changed: initially from a movement aimed at revolutionizing philosophy itself (at least according to some prominent early proponents), into what it is now: an acknowledged, legitimate, but highly specialized subfield within the philosophy of science. (p. 2)

Why did this happen? Bickle identifies two factors. First, the science itself simply advanced too quickly. By a decade or so ago, the real but highly limited understanding of biology conferred by "a standard liberal arts undergraduate education, which most academic philosophers possess, simply was not enough to keep up with the detailed science." He suggests we

> [c]ontrast this situation with state-of-the-art neuroscience circa the early 1980s. Back then, some details of cutting-edge cellular physiology may have required special technical expertise to grasp, but even the Methods sections of experimental publications were written in a tone and at a level of comprehension that

> most college-educated readers could follow. Such is not the case now, and has not been for more than a decade [i.e., since the mid 20-oughts]. (p. 9)

The first factor explains why fewer and fewer philosophers could come and play the revolutionary neurophilosophy game; the second explains why those who did keep playing shifted the scorekeeping steadily away from what philosophers at large would be interested in:

> The field could have remained more central to mainstream philosophy. Its practitioners needed only to eschew the increasing scientific detail and complexity, and instead promulgate more accessible popularized neuroscience toward explicitly philosophical goals and ends. Maybe then even the early hopes of a metaphilosophical revolution would have held sway for longer. But for most neurophilosophers, the increasingly complicated and sophisticated scientific details themselves held greater fascination. The juicy questions lying open at the foundations of the actual neurosciences, not to mention the increasing opportunities to work side-by-side with practicing neuroscientists to address them, trumped the early calls to revolutionize philosophy. Few neurophilosophers seemed to have recognized that they were making any such "choice," at least not explicitly. But the acquired scientific interests of most philosophers who had worked to achieve the necessary background to follow and contribute to actual neuroscience proved to be just too beguiling, not to mention the growing prestige of neuroscience, both within the academy and in the broader intellectual (and commercial) world beyond. (p. 10)

Bickle warns that both of these factors should be expected to influence the ongoing development of experimental philosophy, and the way that experimental philosophers think about their work. And we agree with his observations of them as ongoing trends, which have only continued and strengthened in the few years since he published in 2018. We do not at all advocate trying to fight these trends. Experimental philosophers' increasing technical sophistication and collaborative connections to research in the relevant scientific domains seem to us, indisputably, to be all for the best in terms of the quality of the work being produced. How could anyone argue that the philosophers who are acquiring such mastery, and forging such valuable interdisciplinary links, should somehow be told to stop doing so? We certainly have no interest in making such an argument. But then what are we to do?

The Tools Plus Norms Model

Any useful model for experimental philosophy's future role in the profession must somehow reconcile these two distinct sets of tensions. On the one hand, such a

model must envision the sort of engagement that the universal experimentation model would overzealously mandate—but in a more appropriately limited, targeted way. On the other hand, it must tolerate, and indeed promote, the sort of increasing scientific sophistication that Bickle warns about, and the specialization that comes along with it—but without drifting away from mainland analytic practice and becoming its own disciplinary island. Fortunately for us, analytic philosophy already has a model close at hand with an excellent track record for managing a congruent set of tensions. We'll call this the *tools plus norms model*: we have both a suite of techniques and technologies that are well suited to advancing a set of methodological goals, *and* a corresponding suite of norms prescribing in what circumstances such tools ought to be deployed. The parallel subdiscipline model lacks the force of those crucial norms, whereas the universal experimentation model can be seen as a "tools plus wildly over demanding norms" model. The way to aim for the sweet spot between those two unacceptable models is to construct, implement, and enforce a set of norms of a productively intermediate degree of demandingness.

We do not think we are making up this model of disciplinary uptake of a methodology from scratch, for we find an exemplar in the way that logic and formal methods have been and remain integrated into analytic philosophical culture and practice. Of course, logic and experimental philosophy would be appropriately placed at maximally distant regions of the philosophical methodology manifold on some metrics, most obviously, any metric that weights heavily on "what can be done from the armchair." But, for our purposes, a more useful metric would locate them as very close neighbors, along the lines of "specialized technical tools developed to manage anticipated sources of error," and at some distance from philosophical approaches that are more like, well, just thinking hard about deep stuff.[6]

So what do we learn from the way that logic and formal methods were integrated into philosophical culture and practice? Despite its early sweeping and ambitious phase (there's some reference to this in Bickle's article), by at least the 1980s there was a palpable distinction between two different ways that logic and formal semantics was manifest and valorized in philosophical practice. High-level work by logicians and semanticists was celebrated and seen as a core contribution to the discipline, even when it was on the whole fairly inaccessible to a great majority of working philosophers. Think, for example, about Gödel's incompleteness theorems (Gödel 1931), the Church-Turing Thesis (Church 1936, Turing 1937), Tarski's work on truth (Tarski 1933) and consequence (Tarski 1936), Craig's Interpolation thesis (Craig 1957), Kripke's semantics for modal logic (Kripke 1963), or Cohen (1963) on the independence of the continuity hypothesis and thus the axiom of choice from ZF set theory. While most people with a PhD in

[6] "There is no method except to be very intelligent" T. S. Eliot (1920, 11). We disagree.

philosophy in those days were not expected to know the ins and outs of this work, they were expected to have a high level of comfort with first-order logic, some familiarity with the basics of modal logic, such as the notation itself and some of the properties of S5, and to have at least been force-marched through some of the important early metalogical results like soundness and completeness. And many philosophers, perhaps *most* analytic philosophers, operated somewhere between these two levels, perhaps not as fully licensed, journeyman logicians, but completely capable of applying formal techniques to the kinds of philosophical questions they were interested in answering.

We take it that this picture of formal, technical training is still accurate today, with a functioning consumer-level competence found universally, and many philosophers with at least a bit more than that, and some with much more, all in all integrated with and supported by a comparatively small subcommunity with much deeper technical mastery. There have been some changes; for example, set theory is not as important as it used to be, and decision theory is vastly more important than it once was. But the wide-plains-and-sharp-peaks distribution of skill in logic and other formal techniques still seems to hold. What this exemplifies is the commitment in our professional culture to develop whatever tools we need to do whatever work we want, or perhaps need, to do. Now, it may have seemed that this accelerated stockpiling of mathematical tools was an intra-philosophical augmentation, whereas acquiring the methods of the social sciences has seemed an import from beyond our borders. Nonetheless, we contend that that attitude is merely an unprincipled cognitive atavism caused by our having spent too long in the artificial restrictions of the armchair. It shouldn't matter to philosophical practice whether a proposed new set of tools are armchair-amenable or not; what should matter, in terms of methodological rationality, is whether they will bring us new methodological advantages that are worth the cost.

Crucially, as philosophers have added to our set of tools that *can* be used, we have also developed norms prescribing circumstances under which they *should* be used. Remember, that's the job of the "plus norms" part of "tools plus norms": making sure the tools get used when and where it is of appropriately high value for us to do so. Though typically tacit, such norms are easy to spot when you look for them, and they concern what to do when you reach a point in philosophical inquiry where we can recognize that unaided ordinary language and cognition are too ambiguous, or fallacy-prone, or simply muddled to perform the requisite philosophical labor. This is especially so when there are multiple connectives in play that will need syntactic disambiguation (e.g., "$P \rightarrow Q \rightarrow R$"), or multiple devices that will take scope. In addition to general norms on when formal tools are required, there are subject-specific ones as well, for more specialized sets of formal tools suitable to them. For example, you just can't expect to get too far in the causation literature without being able to deploy and understand neuron diagrams, and maybe also graph theory, the Causal Markov Condition, and so on. And sometimes you might need

to bring out some heavy logical guns and prove some theorems. More often we do something more hybrid, in which the key statements at issue are represented formally, but then the resulting inferences are transparent enough—to those with the right training, at least—that nothing like a lengthy formal derivation is required. Other considerations would include some sensitivity to the strength or surprisingness of the ultimate result, such that the more ambitious the inference, the greater the need for formal clarity of both premises and deduced conclusion.[7]

The fact that formal tools aid most fundamentally with deduction does, however, lead us to offer one substantial caveat about thinking about the tools-plus-norms model for X-phi *too* closely in terms of the model for formal tools. So much of the formal machinery rightfully beloved by philosophers, most especially formal logic, has been finely honed over generations as a way to help us with deductively valid reasoning and thus for the absolute prevention of error. Recall that in Chapter 4 we made a distinction between *P*-strategic disciplines and *S*-strategic ones. In *P*-strategic disciplines, researchers use so purified, but also so *rarified*, a set of methodological resources that they can be extremely confident that there are no errors in their body of results, as with the stockpile of proven theorems in mathematics. We contrasted those kinds of disciplines with *S*-strategic ones, like the empirical sciences, where errors are expected continually to get into the set of considered results, and thus continual attention to the managing and, ideally, removal of errors is required. Of course, significant measures will still be taken by researchers in *S*-strategic disciplines to try to keep errors from creeping into their findings in the first place, but researchers do not merely resign themselves to the imperfections of such preventative cognitive defenses, and to whatever falsehoods may thus get past them: *S-strategies take measures to find them and push them back out the door*.

In these terms, we think that philosophy struggles with the fact that, while it is in practice an *S*-strategic discipline, it has a deeply *P*-strategic self-understanding. We valorize proofs and deductions, and as we have already noted, we train philosophers especially in the use of these kinds of tools and methods and have well-entrenched norms to guide their use. The canonical organizational structure for an analytic philosophy paper is what might be called the *CODA*, or the *compact deductive argument*, with a small enough set of premises and a transparent enough

[7] One might worry that our norms actually err a bit too much on the side of the demand for logical rigor, with the sort of gratuitous formalization that sometimes inflects (and infects) "analyticalese" as parodied by "Why did the philosophical chicken *C* cross the road *R*? To get to the other side Σ." We would note, however, that there are also what might be called pushback norms: ways to reject demands for hyperbolic levels of formalization. For example, if challenged to define a key term more precisely, it is *prima facie* a legitimate strategy to attempt to respond that no further precisification is necessary here for the argument that the term is intended for. Whether, or when, such pushback is successful can itself be contested, of course. We are not looking to argue about any of that here, so much as just to draw attention to these features in contemporary professional argumentation.

inference that readers can see plainly that the conclusion is entailed by them.[8] Correspondingly, the canonical organizational structure for a critical response is to either identify some way, perhaps far-flung, in which the premises do not entail the conclusion, or much more often, and because we are indeed very good at constructing deductively valid arguments, to find some plausible way to reject a premise. And it is right there that the pervasive vulnerabilities to error seep in around the edges: almost always, it seems, some of the premises are either themselves contestable empirical claims on their face or find that their methodological backing can be peeled away and exposed to empirical psychological speculation, as addressed in Chapter 6 with the issue of explaining away verdicts. We can also see now that the problem is even worse than originally canvassed. For it's not just that it is problematically easy to find a rejectable premise here or there for almost any given argument, as our norms for CODA evaluation intensify that problem: typically, you only need to reject *one* premise, in order to thereby reject an argument.

There are places, besides X-phi itself, of course, where philosophers have banged against the methodological walls of this *P*-strategic self-conception. For example, philosophers do sometimes gesture at abductive inferences, especially when they are looking to run roughshod over this or that problematic verdict. But by and large these attempts have not gotten anywhere, not least of all because we simply have no decent norms governing them from the armchair; they become merely a further locus of the kind of dialectical stalemate that we discussed in the previous chapter.[9] What philosophy really needs is a much more comprehensive update of its self-image, in a way that opens us up to some substantial augmentation of its disciplinary norms. And we think experimental philosophy can make a huge contribution. But here is our chief worry: if philosophers are stuck on something like a *P*-strategic conception of philosophical inquiry on the whole, then it can seem that X-phi cannot possibly make the right kinds of contribution to our projects. For X-phi contributions would appear to have to take the form of *establishing* results that can then with maximal safety be relied upon in downstream inquiry, as a sort of empirical verification of an axiom. Yet this is unsatisfactory on the one hand, since the results will still be exposed to nontrivial risk of error, and overly demanding on the other, since it can take quite a long time to fully establish some findings as consensus phenomena. (It is for this reason that we mostly discussed X-phi literatures as *moving us toward* phenomena in the previous chapter, and not so much as having already gotten there.)

[8] In Chapter 2 we talked about this in terms of the tendency of analytic philosophers towards hyperarticulation of premises, presuppositions, and inferential structure.

[9] Weatherson (2003) tries to improve on this methodological state of affairs, but he does not seem to succeed (see, e.g., Weinberg and Crowley 2009). But more to the point, his proposals there do not seem to have caught on at all. See below for further discussion, including more recent proposals by Williamson.

It is thus imperative, when considering what the norms to guide our use of experimental tools ought to be, that philosophers embrace the fact that we work in an *S*-strategic business. We can then propose and properly debate norms that can aid in the more general cause of error management. Importantly, for our purposes here, these kinds of norms can be substantive while still being rather less strenuous than armchair philosophers may fear. Most particularly, if we are right about the requirements of methodological rationality, then it is perfectly kosher for our tools and methods to correct any missteps in the course of inquiry *after* the fact, as part of our ongoing investigative journey. And thus, to head off one concern we have often seen, philosophers do *not* need to sit around waiting for the experimental philosophers to wrap up their studies before getting on with their work. What is important, instead, is that we develop the resources needed to be able, down the line, to get rid of the mistakes we may already have made or will someday make; we do not need these resources to guarantee that no mistakes have been made, past, present, or future. So we are still[10] in enthusiastic agreement with Williamson (2007, 6) when he argues that "it is a fallacy to infer that philosophy can nowhere usefully proceed until the experiments are done." This is right, so long as that philosophical work will over time be in good responsive contact with the kind of inquiry, likely including experimental work, that may help to reveal any errors they may be prey to now without realizing it. And it is to promote the sorts of methodological investments that could make that possible that we now look to articulate in more practical detail just what might be involved in philosophy's adopting the tools-plus-norms model for experimental philosophy.

2 Changing Professional Norms

So, what sorts of tools and norms might we have in mind? In terms of the tools, we already have a small but deep pool of specialists in experimental methods. As Bickle foresees, that trend is likely to persist and intensify on its own. What will likely require some substantial intervention, however, is the continued growth and sustained maintenance of a very broad level of basic experimental competence in the profession at large, especially among the segments that will look to continue using the method of cases and similar techniques. Moreover, we collectively need to devise and implement appropriate norms for the application of such tools. Our ambition here is to do some very preliminary work proposing both policies that could promote and sustain a serious increase in the breadth of competence with

[10] For we would note that this point of concurrence has been an official part of our position for at least fifteen years; see Weinberg (2009, 462).

experimental methods in the profession, and norms to help determine when the use of such tools is to be required. In Chapters 4 and 5 we defended, at a high level of abstraction, the idea that a nontrivial portion of our community's resources should be transferred into error mitigation, including experimental philosophy. We now want to propose a set of specific checks to write against that account. This section will offer some ideas for changes at the level of the infrastructural machinery of the discipline; the next section will zoom down even further, suggesting norms to be followed at the level of individual philosophers.

Before we get into these specific proposals, however, we want to articulate some ground rules for the discussion in order to help clarify what we are trying to do here, and what sorts of responses we most hope to see. We don't want to propose, and do not think it would be helpful to debate, policies that would require large-scale and well-coordinated changes across the profession all at once. Such a degree of explicit coordination is simply not to be expected from analytic philosophy's hundreds of gloriously fractious and strong-willed practitioners. Benefits to philosophical progress may well accrue to the profession globally and over the long term, but whatever actions can be taken to bring these changes about, they will need to be undertaken locally, perhaps by individual members of our community, or in small groups of collaborators and interlocutors, or at most at the level of departmental or editorial practices. We are, thus, restricting ourselves here mostly to ideas that could answer the conditional question: if we can manage in this book to convince just a few handfuls of philosophers about the importance of changing the ways that we approach analytic philosophy, what could they do to start to bring about these changes?

While we do think that some of these proposals are going to seem fairly obvious, others are intended to be conjectural or blue-sky, and we absolutely do not consider these proposals to be even remotely a complete list of possible policies that could, or should, be adopted to improve philosophical practice. Looking forward, then, we very much welcome amendments and novel proposals, especially within the bounds of local implementability just described. By and large this is a level of methodological discussion which has been absent or, at best, *sotto voce*. The vast majority of discussions of philosophical methodology are better understood as either the epistemology or metaphysics of philosophical argumentation. One consequence we hope this book has is to turn up the volume on practical discussions about philosophical methodology. In this spirit, then, it is worth underscoring that these proposals are also, in their way, meant to be *experimental*. We think that philosophers should try them out, see where they work, see where they don't, and revise. Everything we are about to say is fairly speculative about how various changes might play out in the profession. We should be perfectly willing to make adjustments and alterations in light of whatever first-hand discoveries we make and what the practical ramifications of these proposals end up being.

Philosophical Education and Experimental Methodology

With such provisos in mind, let us proceed to the business at hand. We will continue to take the training and cultivation of formal tools as a paradigm for us to learn from, and that paradigm suggests an obvious starting place: graduate education. That is, after all, a huge part of the story about how there is now a broad level of basic competence with logic and formal methods. For generations of analytic philosophers, graduate education has included some significant amount of training in logic and other formal methods. Typically, students are expected to learn the basics of first-order predicate logic, together with some basic set theory and metalogical theorems and some of the basics of modal logic and computability. These are the kinds of things that you get from the classic logic textbooks that have been used to train analytic philosophers. Think, for example, about the kinds of things covered by Benson Mates (1964) *Elementary Logic* and George Boolos and Richard Jeffrey (1974) *Computability and Logic*, or Jon Barwise and John Etchemendy (1999) *Language, Proof, and Logic*, or nowadays by Ted Sider (2010) *Logic for Philosophy*. And students working in technically advanced areas of philosophy that require other kinds of formal tools and methods, for example formal epistemology, would be additionally required to learn the basics of decision theory, inductive logic, Bayesian probability mathematics, and statistics. Think, for example, about the kinds of things covered by Richard Jeffrey (1965) *The Logic of Decision* or Brian Skyrms (1966) *Choice & Chance*, or Colin Howson and Peter Urbach (1989) *Scientific Reasoning: The Bayesian Approach*, or nowadays Michael Titlebaum (2022) *Fundamentals of Bayesian Epistemology*.

This has all been a good thing, done for the good of the profession and the advancement of philosophical inquiry. And we do not think that this kind of graduate education should stop, or that it should be replaced wholesale by training in experimental tools and methods. Different philosophers will need different tools, and even more importantly, we don't want to lose the benefits to the profession that have been produced by inculcating logic and formal tools as part of philosophy's *lingua franca*. We would suggest, instead, a mixed approach, depending on the varying specializations of different departments and the kinds of students that they tend to attract and plan to train. On the model we are proposing, it makes good sense for some departments to center their graduate training on their strong logic programs, especially if they expect to have students who will go on to need high-power versions of these kinds of tools and methods. At the same time, many other departments may offer a more hybrid "tools and techniques" course that might include, in addition to the basics of mathematical and philosophical logic, substantial training in experimental tools and methods. And perhaps some others will find it appropriate to have required experimental methods courses, or perhaps to allow students to take such a course from a social science department to satisfy their degree requirements.

What might this kind of training involve? We can yet again look to the position of formal methods in our field for guidance, for the distribution of levels of competence in formal methods in the discipline also helps answer the question, just how much background in special methods is enough? For both formal and experimental philosophy, the best answer seems to be: a real but modest universal core, then scaling up locally from there as needed. A list of what pretty much everyone in analytic philosophy needs to be comfortable with in terms of logic would include the basic formalism of first-order predicate calculus, the standard suite of rules of inference that go with it and a sense of common sorts of errors with them (as gestured at briefly above, e.g., inadvertently switching quantifier orders), and at least a rudimentary understanding of modal operators. (While almost everyone at least has the main meta-results thrown at them in a graduate logic class, it seems to us that a much smaller percentage are actually later called upon to deploy much understanding of them.) Such knowledge is not just theoretical but also practical; it does you little good to know in principle that affirming the consequent is a fallacy if you're no good at monitoring for them in the wild. From that base, even if you don't plan to be a card-carrying logician, how much mastery of which tools will depend on your interests and the state of the literatures. Compare, for example, how little formal machinery is used in contemporary aesthetics with the amount of formal logic that is used in contemporary metaphysics on topics like identity and grounding (see, e.g., Dorr 2016, Caie et al. 2020, or Cameron 2022). In short: go out and get what you need.[11]

By analogy, let us reaffirm here that we don't even remotely expect all philosophers to become card-carrying experimentalists. Such a state of affairs would not likely lead to a methodologically rational division of labor. Instead, we think that there should be a similar shared core competence in the rudiments of both the psychological foibles that case verdicts may be prone to, and the experimental methods that ought to be used both for investigating and to help compensate for said foibles. This training would inculcate a level of comfort in reading the sorts of psychology and experimental philosophy papers that would be most relevant, including the statistics and methods sections, and a common background knowledge of what sorts of mistakes philosophers need to watch out for when using the method of cases. What we have in mind here is that philosophers who appeal to case verdicts should be aware of any relevant results about cases like the ones that they are interested in, mostly in terms of the domain and the target concept, but also perhaps in terms of other elements of the case. When there are known effects in the vicinity, philosophers should be in a position to take them seriously, for

[11] It is worth mentioning here that there's nothing special about what we are saying here about tools and methods. Historians of ancient philosophy are going to have to learn a lot of Greek, philosophers of art are going to have to be familiar with art history and criticism, philosophers of physics are going to have to know a lot of physics, and so on.

example, by controlling for effects that can be thus controlled. And when there are no already documented effects in the vicinity, appropriately trained philosophers can still keep a savvy eye out for novel effects, and should be able to run very basic sorts of screening tests. Finally, they should be able to recognize when they are in deeper methodological waters than they had meant to wade into, and thus to ask for further help, where needed—one goal of this training for practitioners will be to turn their unknown unknowns of human verdictive cognition, into known unknowns.

We will have much more to say below about what kinds of normative constraints this puts on how philosophers use the method of cases, including what we will call "due diligence screening checks" (along the lines that John Turri (2016) advocates). For right now, we think that it is important to note that the kind of competence that we have in mind can likely be initially developed over the course of a one-semester graduate class like the ones philosophers are now asked to take in order to develop a core shared competence in logic and formal methods. Just like analytic philosophers aren't typically required to master anything so complicated as John Bell's (1985) *Boolean-Valued Models and Independence Proofs in Set Theory* as part of their training in logic and formal methods, so too analytic philosophers wouldn't be required *en masse* to master the methods of linear analysis (say, e.g., Kutner et al.'s (2004), *Applied Linear Statistical Models*) or regression analysis (say, e.g., Fox's (2008) *Applied Regression Analysis and Generalized Linear Models*). What they should be expected to know, however, are the kinds of things that Justin Sytsma and Jonathan Livengood (2016) cover in their *Theory and Practice of Experimental Philosophy*.[12] Sytsma and Livengood provide an introductory guide to both *consuming* and *conducting* empirical research in philosophy, which is meant to familiarize philosophers with the basic set of tools that philosophers can use to design and analyze studies involving surveys about philosophical cases. They walk readers through the basics of experimental design, ranging from basic issues involved with developing research questions and thinking about the strategies that can be used to test hypotheses to how to pick a research design (*simple* or *factorial*, *between-subjects* or *within-subjects*, and so on) and different variations on basic experimental designs (e,g, the use of pretests, blocking, fractional factorial designs, and counterbalancing). They provide a step-by-step guide to conducting empirical research, from basic topics like how to vary the predictor

[12] One aspect of this burden that the experimental philosophy community should usefully shoulder is the development of textbooks and other materials that can be used for this kind of graduate training. We are pointing here to one example, but the analogy to how generations of philosophers have been trained in logic and other formal methods is again helpful—over time high-level experts in these methods wrote a variety of different textbooks, often using different formal systems. (Compare, e.g., Gentzen (1935) or Fitch (1952) with Hilbert and Ackermann (1928).) Wesley Buckwalter's (2024) "A guide to thought experiments in epistemology" is a wonderful example of what we have in mind, covering a variety of errors that experimental philosophers have uncovered and recommending steps that can be taken to mitigate these errors in philosophical practice.

variables and how to measure the response variables to more advanced topics like using R for data visualization and summary statistics and how to analyze parametric, nonparametric, and Bayesian estimation and comparison claims. While they do talk a little bit about more complicated technical machinery, like mediation and regression analyses, the focus throughout is on providing philosophers with the tools that they need to get started.[13]

Of course, in order to implement even the fairly modest but nontrivial changes to how we train graduate students, there will have to be additional investment in experimental philosophy, which will very likely include at least *some* increased hiring of experimental philosophers. Having said that, it is important to note that, when we again look at the development of competency in logic and formal methods across the profession, much of the work was done by a small number of departments who went on to train generations of logicians who went on, themselves, to train generations of other analytic philosophers. Think about the influence that Harvard, Princeton, Stanford, UC-Berkeley, and UCLA, just to name a few of the historically prominent departments in mathematical logic, have had on the development of logical competency in the profession over the past eighty years or so.[14] A lesson to take away from this history is thus, perhaps, that a lot of methodological advantage can be gained for the profession on the whole if we only somewhat increase the number of elite practitioners, just so long as it is accompanied by a fairly widespread augmentation of baseline methodological competence in the profession at large.

Experimental Methodology and Publishing Practices

It is worth pointing out that changes in philosophical publishing practices not only reflected, but also greatly assisted the augmentation in professional competencies in logical and formal methods. The *Journal of Symbolic Logic* was established in 1936 and published many of the most celebrated and important logical

[13] It is worth noting that one substantial positive side-effect of training more of our graduate students in the basics of experimental methods is that this will lower the start-up costs for the increasingly large percentage of students who will go on to become engaged consumers of relevant scientific literatures in their first-order researches, even when this has nothing to do with the method of cases. You don't have to be Alvin Goldman or Hilary Kornblith to think that cognitive psychology will be important to epistemology, and indeed social psychology to social epistemology. And, it is hard to see how conceptual engineering really can work without increasing our understanding of what the psychological impact of using this vs. that concept would be.

[14] There is another important aspect of this analogy to the development of competence in logic and formal methods, namely, the relationships that were formed between philosophy departments and other academic departments where this kind of work was being done—so, for example, mathematics and computer science departments, or linguistics when it comes to formal semantics. We again recommend following this model, and so think that part of implementing the kinds of recommendations that we are making here would involve philosophy departments fostering healthy collaborative relationships with psychology, cognitive science, sociology, and so on.

contributions to our profession, including Kleene's (1938) influential work on ordinals, Church's (1940) and Henkin's (1950) famous work on the theory of types, and Craig's (1953) work on recursive axiomatization, to name just a few early examples. This sort of specialty journal helped foster the community of logicians themselves, while at the same time facilitating the promulgation of their key findings to the greater philosophical community. With this historical example in mind, we further propose that philosophers create a *Journal of Experimental Philosophy*.[15] It would be good to have a place where experimental studies could be reported and discussed without requiring a lot of philosophical stage-setting or lesson-drawing. Philosophical phenomena, like their scientific counterparts, arise from whole literatures and not from individual studies, which was one of the lessons that we learned in the previous chapter. And we have seen too many papers with interesting experimental results that cry out to be published, but nonetheless struggle to find good homes because they do not fit well enough, by themselves, into some specific philosophical literature. A new experimentalist journal could provide a publication home for these kinds of experimental reports where it is hard to fit modest experimental results into the kind of more philosophically ambitious argument frames that would be expected in most mainstream philosophy journals.

This kind of journal could also be home to a specifically empirical version of the "whence and whither" papers, where philosophers take stock of recent developments and trends on a topic. (*Philosophy Compass* papers are often of this sort, for example.) These valuable contributions are more meta than first order, although they often do at least a little bit of stage setting and gesture toward why the relevant literature should move in the philosopher's preferred directions. In this context, we think that this kind of paper would provide a valuable opportunity to take a careful look at a developed or developing philosophical literature, focusing on arguments that seem to turn on verdicts about cases and what the state of local verdictive consensus might be about those cases. One might call it a "verdict audit" of the target literature. Think, for example, about what could be done with the peer disagreement literature in epistemology or trolly-ology in ethics, in terms of taking stock with just where those literatures stand in terms of what is or

[15] This proposal was actually floated among experimental philosophers a few years ago. At that time, we were both opposed to the idea because we were concerned that experimental philosophy had not yet sufficiently established itself within philosophy, and thus posed the risk that experimental philosophers would just end up talking to each other -- which in turn would risk turning experimental philosophy into just another sub-specialty, something we were worried about for the reasons that we've sketched above in section one. And so it is important to us that the mission of any such journal would serve to prevent that from happening. In particular, it should not look to compete with existing philosophy journals to publish the sorts of experimental papers that are already being published in the top generalist philosophy journals. Very interestingly, late in our preparation of this book, we received word that there may indeed be a journal *Experimental Philosophy* being launched soon, under the leadership of Alexander Wiegmann. We very much hope this will be so -- and if so, we hope that some of our suggestions will be found to be useful.

isn't a robust philosophical phenomenon, and moreover, where it may have sorely under-examined empirical commitments.

We have further policy proposals to make regarding philosophical publishing practices. In addition to launching a new journal, we also call for nontrivial changes to how uses of the method of cases should be refereed and, when published, reported in any journal. We have been proposing a modestly nontrivial expansion of the commitment to experimental methods in the profession at large. While some methodological benefits will accrue fairly directly from this investment in the tools, we would conjecture that the real bang for buck comes from developing and enforcing norms mandating the use of those tools in appropriate circumstances, again on a close analogy with the norms requiring the deployment of formal techniques. To this end, we start with the time-honored philosophical practice of making a distinction, and observe that when it comes to enforcing the use of needed tools in philosophy, there are two distinct kinds of norms: norms of argumentation, and norms of publication. In other words, there are two related but nonetheless nonidentical questions to be considered, about the norms governing some corner of professional practice. First: under what conditions will our norms leave philosophers open to an interlocutor's objection, if they fail to use a given tool? Second: under what conditions will a referee be licensed to reject a submitted manuscript, or at least insist on a revision? While very significantly overlapping, these questions can nonetheless receive divergent answers. A proper objection to an argument need not be grounds for revision, should that objection require substantial philosophical work to get up and running; not a reason to reject the paper, but rather a reason perhaps to write a new paper in response. And there are reasons to object to a paper's publication that have nothing to do with whether its arguments are well formed, such as poor citational practices, inadequate engagement with the existing literature, a lack of novelty, or fatal unclarities that render the argument impossible to engage with fruitfully in the first place. We articulate this distinction, because our proposals here will largely concern publication norms, not argumentation norms.

So, while we don't see this so much in philosophy's norms of publication, for obvious reasons, in the empirical sciences an important reason a reviewer may object to the publication of an article is that the paper inadequately reports the experimental methods used in a specific study being reported in the paper or aspects of the data analyses that were used to evaluate the data that was collected. Omissions concerning how participants were recruited, total number of participants, standard measures of effect size, and so on are all grounds to send the manuscript back, even in the absence of a specific reason to think that something has really gone wrong anywhere. Scientists enforce such norms largely for reasons of error mitigation. On the front end, transparency of methods and analyses makes mistakes easier to catch prior to publication, as when a referee notices that an inappropriate statistical test was used. Post-publication, such transparency facilitates

meta-analyses and, of increasing importance, replication attempts. Indeed, the crucible of the replication crisis has sparked the scientific community to forge numerous novel candidate policies, such as pre-registration and open data. And while there are not currently many error-mitigation norms like this in philosophical publishing, there is no principled reason for us not to consider adopting them. Here are three such proposals, which ideally would be incorporated as norms at large in our research community, but fortunately could be implemented at the level of individual editors or even individual referees.

First, while we take the case against the "intuition deniers" to be compelling, for reasons that we discussed in Chapter 2, one of the lessons to be learned from them is that philosophical practice with the method of cases is not as explicit as it needs to be. Philosophers ought to clearly indicate whether their thought experiments are intended to play an evidential role, or are intended to do some other work, such as illustration or clarification. When they are intended as evidential, there should be at least a gloss on what the philosopher's view is as to what grants them positive evidential status, whether it is *qua* folk intuition, an expert judgment, something imported from the background, or whatever. Perhaps most importantly, any supporting text should be clearly tagged as to whether it is intended as an argument for the case verdict, or to help draw the reader's attention non-argumentatively to key features of the case, or yet some other purpose.

We should be clear, though, that we are *not* saying that philosophers need to *defend* their disclosures. If they say that they take themselves to be offering an expert judgment, they do not thereby owe us any sort of argument on behalf of experts in the relevant domain. Other features of specific arguments may incur such an argumentative debt, but that is outside the scope of our considerations here, where we are recommending a norm for publication. This is a norm about transparency and bookkeeping, not one about preemptively securing the truth of a case verdict. The point is just that philosophers should make very clear just what they have in mind when they deploy the method of cases, since we now know that there are problematic ambiguities about the thought experiments in many philosophical texts. As the scientists do, so too should philosophers require one another to be transparent about their methods.

Second, philosophers deploying an evidential case verdict should include a brief, maybe *very* brief, literature review of relevant results, of course including not just worrisome findings but also ones suggesting that the case verdicts in question should be expected to be robust. Just how long such a review will be depends, of course, on the state of play in the relevant experimental literature, but we really do not mean for it to impose a giant tax on everyone's article word-budget. The point is just to make sure people are taking stock of what work has been done. For example, it should also be fine to just piggyback on already published reviews, and where appropriate, simply cite one such extant review. Often the results surveyed in such a review will be *prima facie* unproblematic, but when they are not,

philosophers will need to decide how much of a defense of their use of the case is needed, and referees and editors can decide in turn whether the authors addressed the issue adequately. And, when a literature review reveals documented vulnerabilities, it will also hopefully include ways in which our practices can mitigate these errors. Again, Buckwalter (2014b) provides a wonderful example of just the kind of thing that philosophers should be consulting when performing a literature review.

Third, when philosophers are interested in using case verdicts that have *not* already been somewhat well studied, then they should *either* at a minimum run a due diligence screening check (see below) *or* at least offer a brief explanation as to why they think none is necessary. Now, the second disjunct might seem to present a loophole so large as to debilitate the norm on the whole. After all, if philosophers can always just say, for example, "as a native speaker, I feel my command of the language is adequate to know English speakers would have a strong tendency to categorize that as a case of knowledge," and thereby check the box and move on, why would anyone go through the hassle of running an actual study? Our answer is fourfold. First, simply needing to put a sentence like that to paper—permanently, publicly, in print—will bring many philosophers up short, and cause them to ask themselves whether they really *do* believe in their command *that* strongly. Second, these kinds of declarations paint nice, clear, high-value targets for those who are willing and able to run the studies. Philosophers looking to make this side-step move to avoid doing any studies will thus have to weigh the risks of someone else circulating the rather embarrassing discovery that their nose for the lingo is not as discerning as they had proclaimed it to the world to be. Third, as the base-level competence in experimental techniques grows, it will become much less burdensome to run such studies in the first place. We trust that many philosophers will come to see the value of such empirical work, and be happy to contribute to it when it is fairly straightforward for them to do so. And sometimes a case really will be so well attested in ordinary discourse that such an attitude of confidence is fully legitimate, and the risk of some future public empirical embarrassment sufficiently remote to be a rational bet for philosophers to take. (Which is not to say it wouldn't *also* be worth it for some empirically minded philosopher to include the scenario in a future study, just in case!)

We want to re-emphasize that these proposed norms are all about error mitigation, and *not* meant as forms of argumentative objections, and especially therefore are *not* to be assimilated to the model of securing a premise or plugging an enthymeme. In particular, *nothing at all* in the above should be construed as requiring either that philosophers compellingly demonstrate the legitimacy of their preferred version of the method of cases before deploying it; or that they do anything like antecedently establish empirically that the case verdict that they wish to appeal to is robust. Rather, these requirements would serve to render visible the empirical commitments incurred by philosophers as they use these methods. Often

these commitments will be taken up on spec, and not pre-paid—a situation that is only methodologically rational once a healthy broad competence in experimental methods has begun to be cultivated. They should constitute fair targets for future challenges from other philosophers in other papers, but ought not be treated for purposes of publication as undefended premises and criticizable as such.

3 The Post-Armchair Method of Cases

Our proposals for modifications to analytic philosophy's training and publication practices aim to better integrate experimental philosophy into ordinary philosophical practice, in order to better enable the sort of error detection and correction capacities requisite to put our community back into a state of methodological rationality. While we have tried to make our initial proposals fairly concrete, they still largely apply at something of a community level. But we suspect that many philosophers, including many who are sympathetic to our case, will still want to know what this all means in terms of the individual professional lives of philosophers—what it means, most specifically, to their own philosophical practices. In this section, we will try to spell that out in practical terms: just what are we advocating that analytic philosophers should *do*, in their day-to-day practice, if they intend to continue using the method of cases in a methodologically rational way?

Before we start, it is important to again underscore the fact that many core aspects of philosophical practice, including many ins and outs of the method of cases, will remain the same in our proposal. Philosophers will still find themselves contemplating verdicts where those verdicts put pressure on some philosophical theory and could be better accommodated by another, perhaps only slightly different one. Much of our disciplinary training, our linguistic and theoretical mastery, our committed concern for conceptual niceties, our finely honed sense of where two deeply similar philosophical theories may come apart from one another, and so on will still come very much into play in daily philosophical practice. There is simply no sense whatsoever in which our proposal involves turning over philosophical questions wholesale to empirical scientists. The scientific tools and methods that we have in mind are not applied *in loco philosophorum*, but rather, *in philosophico situ*. This is, perhaps, the most important lesson that we can learn from looking at the way that logic and formal methods are used in contemporary analytic philosophy: without needing anything so lavishly expensive as a norm of universal formalization, the analytic philosophy community bought itself a truckload of error management for the substantive, but by no means exorbitant, cost of installing logic and other formal tools in its workshop, along with some moderate norms for their use. We think that similar terms and conditions can be used for the installation and deployment of experimental tools here, as well.

Keeping this in mind, the first norm for responsible practitioners of the method of cases would be: *use experimental tools and methods in situations where we expect armchair resources to pose an empirically live risk of falling short*. Such a norm would operate on close analogy with our norms for requiring the use of formal tools and methods. As we noted above, while we do not require across-the-board formal derivations, we do possess a reasonably good sense about what sorts of inferential steps may be so tricky that we ought to give them a more mathematical treatment. A paper claiming a non-obvious entailment from even a small but moderately complicated set of propositions with, say, iterated modal operators, or even just a handful of nested quantifiers of first-order logic, will likely be required to include at least enough machinery of proof to make the entailment perspicuous. In the same spirit, philosophers should lean heavily on experimental tools and methods to learn where unaided human cognition is unacceptably susceptible to error when engaging in philosophical argumentation. And, for a substantial number of these potential foibles, good experimental tools and methods can help provide excellent resources for overcoming them. For example, while armchair philosophy offers practically no resources for addressing error vectors like order effects, or sample bias, especially when intensified by motivated cognition, it is nonetheless very easy to control for order in an experimental design, and outliers can similarly be easily detected, so long as samples are gathered well. Some cognitive limitations, such as motivated cognition and related phenomena like myside bias, have proved incredibly thorny problems (for recent discussions, see Levy 2021, Mercier and Sperber 2017, and Stanovich 2021). Probably nothing could do so, and indeed the total removal of any source of error is probably a Cartesianly unachievable demand. But there are still significant benefits to incorporating experimental tools and methods in these cases. In this particular spot, mitigating the threat of motivated cognition would be a significant benefit, even if it falls unfortunately far short of totally eliminating it.

It is important to keep in mind an important disanalogy between norms for formal tools and those for experimental ones: although we can now take ourselves to have at least *some* good ideas about what sorts of circumstances, materials, and so on will render case verdicts vulnerable to error, we also should recognize that there is much, *much* more that we just don't know. As we argued in Chapter 2, our priors at this point should be that we should expect a fairly broad baseline of stability, against the backdrop of which we should predict the emergence of unpredictable failures of robustness in the case verdict evidence. As we learn more—hopefully quickly, due to increased investment in experimental philosophy!—our norms for individual practice will have to change, as well, and in ways that we are sketching here. But it is up to the community on the whole, drawing also on the work of scientific psychologists and with perhaps a special burden falling on X-phi specialists, to steadily uncover the error profile of the human philosophical instrument. And as that error profile gets steadily more filled in, we will be able as

a community to have much better-informed guidance as to where, when, and how experimentation ought to be required. Some of these problems may be addressable from the armchair, such as the radical step of simply foregoing using a kind of case for which there is ample evidence of vulnerabilities. But one thing we know now about many of these risks of error, even at this early stage, is that good experimental tools and methods could help provide excellent resources for overcoming them. And that, therefore, the norms will recommend or even mandate the deployment of such methods.

The error profile will include specifications both general and local. There will be lore about what kinds of errors may apply fairly widely, but also more targeted findings such that philosophers looking to make use of a specific philosophical case, or set of cases, should be expected to find out what the literature may already say about them. If the publishing norm we suggested above is widely adopted, then this kind of background work will be mandatory. But even in the case that this publishing norm is not widely adopted, checking out the existing X-phi work on a case is obviously a good idea. If there are studies on highly similar cases, then the approach that philosophers take to their novel case should be informed by what these studies have revealed about similar cases.[16] For example, epistemologists looking to use Gettier cases may want to read up a bit on which ones seem to work widely with the folk and which ones don't (see, e.g., Starmans and Friedman 2012 and Turri et al. 2016, although see also Gonnerman et al. 2022), and which ones seem to be potentially sensitive to things like order effects (see, e.g., Machery et al. 2018). Often philosophers may find that the verdicts they are looking to discuss have been well examined and can be expected to be unproblematic, but it would be good to know whether, on the other hand, they are ones that have been found pretty much only among philosophers, and perhaps are not to be trusted as robust.[17]

Let's consider three different strengths of worry that might come into play for a would-be verdict-deploying philosopher in the post-armchair era, who is looking to appeal to some verdict about a particular case. First, what should they do if there is some literature on that case or highly similar ones, and it is still in an unsettled and at least somewhat discomfiting state? Second, what should they do if the literature is converging on the view that the specific case is, in fact, highly nonrobust? Third, what should they do if the literature appears to be converging on the view that none of the cases in a given domain are passably robust? We think these

[16] It is important, of course, to keep in mind that what counts as a relevant case may not be something that can be read right off of the substance of a given case. Consider the discussion in Chapter 2 concerning Wiegmann and Waldmann's (2014) hypothesis about order effects in trolley cases. As we indicated there, the mechanism of order effects that they consider may extend well beyond the end of the trolley line. This means that it might help for analytic philosophers to run their novel cases past an experimental philosopher who could perhaps recognize any such possibilities based on the developing state of the literature.

[17] See our discussion of stakes effects in Chapter 6; or also, Starmans and Friedman (2020), on the sometimes highly eccentric knowledge attributions of philosophers, even compared to other academics.

questions cannot really be answered too definitively in advance, and will likely depend on highly specific shapes of the relevant literatures at a time. Nonetheless, we think it would be useful to the sympathetic but nervous practitioner of the method of cases if we can give them a bit more of an idea of what we would have in mind for them should they encounter experimentally adverse circumstances.

For the first sort of circumstances, where no clear picture about the case has yet emerged, we would recommend that the philosopher proceed with caution, and eyes wide open, recognizing that their paper may be vulnerable to later being shown to be based on a piece of putative evidence that is in fact spurious. They should also, hopefully, consider what activities they can engage with to improve our evidentiary picture of the target case—practice what we are calling "defensive x-phi" below. Are there maybe other cases they could look to in order to make their argument? Could they perhaps run, or collaborate with an experimental philosopher (who their departments have hopefully already hired; see above), to run some further X-phi on the case to try to help resolve its status? Specific experimental remedies may also be available: as we have noted, if a case fundamentally suffers from order effects, then they could look to run a study that would average responses across a range of different orders, with the hope that this averaged result will be sufficiently conclusive to allow them to put it to work.

For the second sort of circumstances, where we have good empirical reason to positively expect some individual case to be broken, here we are inclined to advise the philosopher to pursue other argumentative avenues. We noted above that the "fake barn" cases have just over and over again failed to give results like those called for in the epistemological literature where they originated; they seem to average out somewhere a bit, but just a bit, on the "it's knowledge" side of the line. Philosophers should probably just not look to use that case as an instance of non-knowledge anymore. Suppose a theory of knowledge A can only gain traction over a rival theory B by means of A's alleged better handling of fake barn cases as not known. In that case, philosophers need to acknowledge that they may just not have a reason to prefer A to B; indeed, they may have reason now to prefer B over A. We suspect that in a real version of such closely matched rival theories, reaching past the method of cases will often provide a way of moving the debate forward, and should this not be so, then we may have to settle for the very modestly skeptical result that the human philosophical instrument does not have quite the resolution to prefer one over the other. If A and B agree on literally every other case out there, then we perhaps should just accept that if "A or B" is the best we can do in epistemology, it's still pretty good.

But what about the third scenario, where an entire domain of verdicts seems non-robust?[18] Should we generalize from the preceding paragraph, and tell all

[18] We thank Jessica Brown for pressing us to address this worry.

those philosophers to just find a different area to work on? While this does seem a potentially worrisome outcome, our first response is to consider it sufficiently far-fetched that it should not keep either us or the method of cases practitioners up at night. Again, for many of the problems that X-phi uncovers, more X-phi can help address those problems, as with averaging across orders, or determining that some subset of cases produce culturally universal (or near-so) verdicts, or by providing modes of theorizing that can make good use of inconclusive verdicts. Nonetheless, we feel inclined to accept the worst-case scenario here—that X-phi will reveal a whole domain of case verdicts may be so awash in noise, so attenuated in actual signal, that there is just nothing to be done about it—is indeed *a very, very bad case*. Should such a scenario actually obtain, it may well be that the method of cases for that domain should be declared evidentially bankrupt: their methodological house is in such poor shape that X-phi cannot help to renovate it, but rather helps us see it must be condemned. We want to underscore that we see this as neither a likely nor a desirable result. But if we find ourselves in such a situation, the right thing to do is not to push on with these hopelessly flawed cases, but rather to look for radically different methods than case-based ones.

Those are all situations in which a philosopher has some degree of evidence ready at hand about a particular case or set of cases. But what should individual philosophers do when they want to put significant argumentative weight on a specific verdict about a novel case or cases, that is, ones that have not yet been empirically investigated *at all*, and so where they don't already have specific information about how our verdicts about such cases might run into trouble? This kind of situation requires a second individual-level norm: *in the absence of specific information, practice defensive X-phi*. Using such a case will of course start with the sort of literature review already discussed; otherwise, how will the philosopher know that it is, in this sense, a novel case? Once that review has been completed, and no specific concerns have been thereby revealed, philosophers should be prepared to run a *due diligence screening check* for novel cases: they should do some very preliminary, and for that matter even fairly superficial, work just to check for the most common sorts of error vectors. And, since we expect that the kinds of changes in graduate education will lower the costs associated with this kind of experimental work, we think that it would thus be reasonable, and not unduly burdensome, for them to run a few different variations on any novel case, considered in different orders against perhaps a standardized set of anchor cases, and checked across a reasonably diverse pool of participants. This would not require anything complicated in terms of statistics, but would instead involve using the kinds of low-intensity experimental tools that would be a standard part of revised graduate education, as proposed above (and easily enough acquired if those proposals are not implemented!). While such quick due-diligence checks would not be taken as any sort of definitive demonstration that the desired verdict was the verdict about the case, nonetheless philosophers performing these kinds of due-diligence checks could

rightly feel increased confidence that their work was shielded from some of the more common sorts of errors.[19]

Our focus throughout this book has been on the question of what philosophers can do to help shield the method of cases from error, both in terms of keeping those errors from arising in the first place, but also in terms of weeding them out when they pop up. In Chapter 3, we talked about the fact that some modes of inference are error-fragile, and others error-robust. Error-fragile modes of inference are very likely to yield false conclusions when even a small number of premises are wrong, in fact, for really error-fragile modes of inference, one false premise can be enough to send things wildly off-track. By contrast, error-robust modes of inference can withstand a nontrivial number of false premises and still lead to true conclusions. Our point of bringing out this distinction in Chapter 3 was to observe that the standard analytic practices of theory selection tend to be significantly error-fragile, and thus even a small risk of error in the verdicts that make up the premise set entails a much more significant risk that our philosophical theories are false. As we discussed there, this is one of the reasons why mere reliability of the armchair method of cases is not enough to establish the method's trustworthiness. And, as we discussed in Chapter 4, since philosophers are using such an error-fragile norm of theory selection, then we need to invest heavily in detecting, catching, and even quarantining errors in their case verdict evidence on pain of methodological irrationality. But there's another option for error-management: broadening our inferential toolkit!

This suggests a third individual-level norm: *deploy more error-robust modes of inference and theory selection*. Deductive argumentative norms have long had pride of place in analytic methodology. But once we more fully embrace the methodological fact that our approaches to error-management must be *S*-strategic whether we like it or not, whole new directions of methodological inquiry and innovation can open up. And while most of the metaphilosophical literature in the "armchair versus X-phi" debate has focused on the evidential status of the case verdict data, there are at least two other kinds of methodological questions that can be asked once we realize that the issue isn't just about our data, but about our *methods* and *practices* using that data—a point that we emphasized in Chapter 2. The first question is what sorts of non-monotonic, error-robust modes of inference could we pursue? And the second question is what kinds of more *exception-tolerant* forms of philosophical theories can we look to establish using those modes of inference?[20]

One highly suggestive avenue to pursue here is the class of more soft-edged variants on universal generalizations. For reasons we saw in our discussion of Jennifer Nado's work in Chapter 3, when we pursue generalizations like *All Fs are Gs* on case

[19] And, in cases where these preliminary studies revealed problems, analytic philosophers could team up with experimental philosophers to perform more careful experiments involving the cases.

[20] See Weinberg (2016b). Thanks to Wesley Buckwalter for this useful piece of terminology.

verdict data, we make inquiry disturbingly sensitive to even very slight amounts of noise in that data. For strong theoretical claims like universal generalizations, one or two blown cases can shipwreck inquiry. (Suppose, e.g., that we have a case where it appears falsely that *some F is a G* when it is not and that case is used as partial evidence that the universal generalization is true, or perhaps worse, a case where it appears falsely that *some F is not G* and that case is used as evidence that the universal generalization is false.) We've focused most of our attention on being able to catch problematic cases like this as they emerge or soon after. But there are also highly meaningful, substantive *generalizations* that are just not so easily sunk, such as

Most F's are Gs
F's are typically Gs
F's are normally Gs
Ceteris paribus, Fs are Gs
Fs are Gs (understood as generics)[21]

We don't think that there's any mystery why such exception-tolerant generalizations have not yet been popular in analytic philosophy. From the armchair, such generalizations are both hard to infer to and hard to infer from. On the 'inferring-to' side, they don't seem like the sorts of theoretical claims that could be properly vetted from the armchair and whatever cases we might choose to consider from a reclined perspective. Just consider the easiest version on the list, *most Fs are Gs*. If we come up with a bunch of cases of *F*s, and for even, say, 80% of them we render the further verdict that they are *G*, we need some sort of reassurance that we are not cherry-picking. As we argued in Chapter 3, even if we are confident that each of our individual *F*s are typical ones, that does not mean that our overall sample is representative of *F*s on the whole. On the 'inferring-from' side, just consider how *all Fs are Gs* and *all Gs are Hs* straightforwardly entails *all Fs are Hs*. Replace the "all"s with "most"s and you have a fallacy. Such claims are not as amenable to being deployed as premises for further argument, if we insist on such arguments being held to the exceptionless standards of logical validity. But what makes such universals attractive is the same feature that makes them dangerous to use with noisy data: their inferential strength comes with the inextricable cost of great error-fragility.

But once the profession has invested more substantially in promulgating more experimental tools, these concerns about how to infer both to and from soft

[21] Adapting our inferential practices in these sorts of ways is a far from trivial matter, metaphilosophically speaking. See for example Johnston and Leslie (2012) or Sterken (2016), for some ideas about how tangled the connections might be between generics, philosophical analyses, and predications. But it is nonetheless a direction of progress worth exploring.

generalizations should weaken and fade. Just to sketch one example, one way of testing a claim that *Fs are normally Gs* would be to take any putative case of an F which is not G and present that vignette to a panel of participants who will be asked how unusual or weird or abnormal the situation is. More generally, since the inductive and abductive methods of the empirical sciences have proved capable of both establishing these sorts of claims, and then using them as a basis for further inquiry, perhaps as phenomena to be explained, there's no reason to judge that such methods won't eventually work well for philosophy in the same way. And we think that there could be an early heyday here of rehabilitating generalizations which have failed as universals, but which may well hold up as informative and interesting soft-edged ones. Surely *something* interesting would follow if, *typically* when we know that p, we are in a position to know that we know p? Or if the identity between knowledge and justified true belief holds after all, even if only *ceteris paribus*?

Such suggestions are, we think, low-hanging fruit here, easily grasped and enjoyed from a standing position. A further bounty of error-robust modes of inference come along when we shift from thinking about philosophical *theories* toward thinking about philosophical *models*. With rival theories, at most one can be true, and as we have seen, often argumentative methods are not up to the task of securing which one (if any) of a set of competitors is correct. But with models, we can take measures both of how well a given model fits the phenomena, and how to penalize models for introducing added complexity in order to chase a closer fit.[22] If we can really get ourselves to a point where we are building quantitative models, that would unlock a whole new set of tools for helping us secure more consensus on how such trade-offs should go. A lot of other methodological changes can come along with shifting to thinking in terms of models. One major shift from thinking in terms of theories to thinking in terms of models is that scientists take it as a given that all models are, strictly speaking, false: we expect them to involve various simplifications and idealizations; any given model is aimed at capturing some set of the phenomena and perhaps simply not attempting others; some parameters will correspond to actual features of reality, and others will be there just to make the math work out.[23] Thus another form of progress here might be allowing for a pluralistic proliferation of models, the successful ones each capturing some key piece

[22] Some philosophers already treat philosophical theories like models in the sense that we have in mind here. So, for example, Lewis (1983, 353) writes about the metaphilosophy of measuring theoretical costs, "What's true is that a theory may be faulted for its overabundant primitive predictions, or for unduly mysterious ones, or for unduly complicated ones. These are not fatal faults, however. They are to be counted against a theory, along with its faults of overly generous ontology or of disagreement with less-than-Moorean commonsensical opinions. Rival philosophical theories have their prices, which we seek to measure. But it's all too clear that for philosophers, at least, there ain't no such thing as a free lunch." And, Richard Boyd (1990) defends scientific realism by comparing the costs and benefits of what he calls the realist "philosophical package" with those of the empiricist package and the constructivist package.

[23] See Godfrey-Smith (2006) and Weisberg (2007).

of philosophical phenomena, instead of a *Highlander*-style "there can be only one" deathmatch between philosophical theories, which often tends in practice more towards a *Reservoir Dogs*-style "there won't even be one."

Indeed, one further consequence of adopting the perspective that we are proposing here is that we may well be able to feel legitimately like we are making more progress on philosophical questions even when we aren't yet arriving at a univocal answer to them, so long as we are steadily accumulating phenomena. This is a kind of answer to the pessimistic induction: don't worry if theories come and theories go, and maybe it's really appropriate to adopt an attitude of moderate skepticism toward the theories that our current time-slice of history has set before us. Nonetheless, we can see philosophy as making steady progress if we measure it by the accumulation of phenomena, not the triumph of any particular theories, so long as addition of new phenomena also steadily leads to new theories that can accommodate them well. This is an idea that has been given some slogan-level endorsement by some very prominent figures, but working out the norms to guide such inferences in actual philosophical practice is a task for the analytic philosophical community in, we hope, the near future.[24]

4 Experimental Philosophy Is Philosophy-Loving Philosophy

We are obviously only sketching *some* possible norms here in broad outline, and we are sure that there should be others to consider, both along these lines but either more or less demanding, or concerning other aspects of the profession altogether (such as, say, norms of authorship credit). We hope to have made two points clear by this exercise, however. First, it is helpful to see that there are a range of methodological norms that could be adopted by our profession in order to reap the benefits of experimental tools and methods but without anything like the counterproductive extravagance of the universal experimentation model. We think that much of the resistance to experimental philosophy originates in a fear that experimental philosophers must have something just that crazy in mind. What we have been trying to stress throughout the previous two chapters is that we could purchase a sizable, pleasantly inhabitable patch of methodological improvement at what is really a rather low professional cost. Second, although the norms sketched above are all fairly modest, they nonetheless overthrow any conception of philosophy as an armchair discipline. Adoption of even these modest norms of empirical and experimental integration would represent a significant departure from armchair philosophy, conceived at the level of the profession on the whole. Any

[24] E.g. Paul (2012), Williamson (2017). See also Weinberg (2016b, 2017) for some very preliminary attempts to draw methodological lessons from thinking about philosophical inferences in terms of model selection.

particular armchair philosopher may perhaps be licensed to remain thus seated, but at best for only so long as they are in the right kind of responsive contact with a whole bunch of other philosophers who are not. Retaining a minority of armchair *practitioners* can only be methodologically rational so long as those remaining armchairs are integrated into what is overall an empirically informed and empirically guided *set of practices*. Some individual philosophers can pretty much restrict themselves to armchair philosophy, so long as philosophy on the whole does not, and so long as our methodological resources expand in ways that facilitate our detection of, and compensation for, the sorts of errors that these resolutely armchair philosophers may be unknowingly prey to.

At this point it is worth going back to the beginning to reflect again on one of the common responses to experimental philosophy, namely, that it results in a kind of skepticism or a kind of philosophy-hating scientism. We hope that this book demonstrates that these charges are inaccurate. As we have shown, we do not need any thesis to the effect that case verdicts are on the whole evidentially bankrupt in order to raise worries about how analytic philosophers use the method of cases in ordinary philosophical practice. It is more than enough to point out that our armchair philosophical resources are inadequate for telling us much about when and where the method of cases may go off the track. This is not skepticism, but just a realistic and informed view of the limits of armchair philosophy. And we are not imposing any sort of unreasonably high demands on our use of the method of cases that would fail on analogy with our ordinary epistemic lives or with the sciences: our ordinary epistemic lives are rightly held on the whole to a lower standard, and the sciences generally can meet those standards where armchair philosophical practices cannot, not the least of all because the sciences make ample use of the sorts of tools that cannot be accessed from the armchair.

As for the charge of scientistic philosophy-phobia, it is important to stress that criticizing philosophical practice for susceptibilities to error, and advocating reforms, is a time-honored philosophical activity in its own right, practiced by such luminary figures as Descartes, Hume, Kant, the pragmatists, the logical positivists, and Austin. We are, of course, not claiming that any work by experimental philosophers makes us into Humes reborn. Our point is simply that offering these sorts of arguments in no way entails either scientism or being a "philosophy-hating philosopher." We are not at all calling for the dismantling of philosophy to be replaced by some scientific ersatz. We are calling instead for constructive revisions to, and practical debates about, current philosophical practice.

There is perhaps some interesting similarity between experimental philosophy and the way that Descartes deploys hyperbolic skepticism as a tool for inquiry. Both are motivated by the discovery that our cognitive lives are more shot through with error than we had expected. Both strive to identify the worrisome cognitions, and then make a novel application of philosophical tools to restore confidence to those that deserve it, and prevent further mischief from those that do not. But they

diverge sharply regarding the standards for flagging something as methodologically worrisome. For Descartes, famously, the mere conceivable possibility of error is enough for the suspension of belief. For experimental philosophers, error possibilities must be live and empirically attested. As a consequence, Descartes ultimately did not really have the resources he needed in order to pull back from his nosedive of doubt, with too much called into question and not enough resources left over to recall them to trustworthiness. The kind of methodological challenge that we have raised here faces no such worry since the sorts of concerns that its empirical methods raise can be addressed by those same methods. This is the reason that we suggested at the very beginning of the book that philosophers should opt for an *unbounded* conception of philosophical methodology rather than a *bounded* conception of philosophy's ambitions. Of course, this will involve changing both philosophy and philosophical practice, but as we have shown, the changes are less radical than philosophers sometimes worry and the benefits we get from changing the character and shape of philosophical practice outweigh the costs associated with business as usual.

References

Alexander, Joshua. (2012). *Experimental Philosophy: An Introduction*. Polity.

Alexander, Joshua. (2016). Thought experiments, mental models, and experimental philosophy. In Jennifer Nado (ed.), *Advances in Experimental Philosophy and Philosophical Methodology*. Bloomsbury Press, pp. 53–68.

Alexander, Joshua, Diana Betz, Chad Gonnerman, and John Waterman. (2018). Framing how we think about disagreement. *Philosophical Studies* 175: 2539–2566.

Alexander, Joshua, Chad Gonnerman, and John Waterman. (2014). Salience and epistemic egocentrism: An empirical study. In James Beebe (ed.), *Advances in Experimental Epistemology*. Bloomsbury, pp. 97–118.

Alexander, Joshua, Ron Mallon, and Jonathan M. Weinberg. (2010). Accentuate the negative. *Review of Philosophy and Psychology* 1: 297–314.

Alexander, Joshua, and Jonathan M. Weinberg. (2007). Analytic epistemology and experimental philosophy. *Philosophy Compass* 2: 56–80.

Alexander, Joshua, and Jonathan M. Weinberg. (2014). The "unreliability" of epistemic intuitions. In Edouard Machery and Elizabeth O'Neill (eds.), *Current Controversies in Experimental Philosophy*. Routledge, pp. 128–145.

Andow, James. (2015). How "intuition" exploded. *Metaphilosophy* 46: 189–212.

Angelucci, Adriano. (2022). On justifying case verdicts: A dialectical hypothesis. *Inquiry*. 10.1080/0020174x.2022.2092905

Appiah, Kwame Anthony. (2008). *Experiments in Ethics*. Harvard University Press.

Austin, John. (1956). A plea for excuses: The Presidential address. *Proceedings of the Aristotelian Society* 57: 1–30.

Babcock, Linda, and George Loewenstein. (1997). Explaining bargaining impasse: The role of self-serving biases. *The Journal of Economic Perspectives* 11: 109–126.

Ballantyne, Nathan, Jared Celniker, and David Dunning. (2024). Do your own research. *Social Epistemology* 38: 302–317.

Barwise, Jon, and John Etchemendy. (1999). *Language, Proof, and Logic*. CSLI Publications.

Baz, Avner. (2017). *The Crisis of Method in Analytic Philosophy*. Oxford University Press.

Baz, Avner. (2023). On philosophical idling: The ordinary language critique of the philosophical method of cases. *Synthese* 201: 1–20.

Bealer, George. (1996). A priori knowledge and the scope of philosophy. *Philosophical Studies* 81: 121–142.

Bealer, George. (1998). Intuition and the autonomy of philosophy. In Michael DePaul and William Ramsey (eds.), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Rowman & Littlefield, pp. 201–240.

Bealer, George. (2004). The origins of modal error. *Dialectica* 58: 11–42.

Beardsley, Monroe. (1982). *The Aesthetic Point of View: Selected essays*. Cornell University Press.

Beebe, James, and Wesley Buckwalter. (2010). The epistemic side-effect effect. *Mind & Language* 25: 474–498.

Beebe, James, and Ryan Undercoffer. (2015). Moral valence and semantic intuitions. *Erkenntnis* 80: 445–466.

Beebe, James, and Ryan Undercoffer. (2016). Individual and cross-cultural differences in semantic intuitions: New experimental findings. *Journal of Cognition and Culture* 16: 322–357.

Bell, John. (1985). *Boolean-Valued Models and Independence Proofs in Set Theory*. Oxford University Press.

Bengson, John. (2013). Experimental attacks on intuitions and answers. *Philosophy and Phenomenological Research* 86: 495–532.

Bengson, John. (2014). How philosophers use intuitions and "intuition". *Philosophical Studies* 171: 555–576.

Bengson, John, Terence Cueno, and Russ Shafer-Landau. (2022). *Philosophical Methodology: From Data to Theory*. Oxford University Press.

Bergenholtz, Carsten, Jacob Busch, and Sara Kier Praëm. (2021). Further insights on fake-barn cases and intuition variation. *Episteme* 20: 163–180.

Bickle, John. (2019). Lessons for experimental philosophy from the rise and "fall" of neurophilosophy. *Philosophical Psychology* 32: 1–22.

Blok, Sergey, George Newman, and Lance Rips. (2005). Individuals and their concepts. In Woo-Kyoung Ahn, Robert Goldstone, Bradley Love, Arthur Markman, and Philip Wolff (eds.), *Categorization Inside and Outside the Laboratory: Essays in Honor of Douglas L Medin*. Washington, D.C.: American Psychological Association, pp. 127–149.

Blouw, Peter, Wesley Buckwalter, and John Turri. (2018). Gettier cases: A taxonomy. In Rodrigo Borges, Claudio de Almeida, and Peter Klein (eds.), *Explaining Knowledge: New Essays on the Gettier Problem*. Oxford University Press, pp. 242–252.

Bogen, James, and James Woodward. (1988). Saving the phenomena. *Philosophical Review* 97: 303–352.

BonJour, Laurence. (1985). *The Structure of Empirical Knowledge*. Harvard University Press.

BonJour, Laurence. (1998). *In Defense of Pure Reason*. Cambridge University Press.

Boolos, George, and Richard Jeffrey. (1974). *Computability and Logic*. Cambridge University Press.

Bourget, David, and David Chalmers. (2014). What do philosophers believe? *Philosophical Studies* 170: 465–500.

Boyd, Richard. (1990). Realism, approximate truth, and philosophical method. In Wade Savage (ed.), *Scientific Theories – Minnesota Studies in the Philosophy of Science, Volume 14*. University of Minnesota Press, pp. 355–391.

Brown, James. (1994). *Smoke and Mirrors: How Science Reflects Reality*. Routledge.

Brown, Jessica. (2013). Intuitions, evidence, and hopefulness. *Synthese* 190: 2021–2046.

Brown, Jessica. (2017). The Gettier case and intuition. In Rodrigo Borges, Caludio de Almeida, and Peter D. Klein (eds.), *Explaining Knowledge: New Essays on the Gettier Problem*. Oxford University Press, pp. 191–212.

Buckwalter, Wesley. (2010). Knowledge isn't closed on Sundays. *Review of Philosophy and Psychology* 1: 395–406.

Buckwalter, Wesley. (2014a). The mystery of stakes and error in ascriber intuitions. In James Beebe (ed.), *Advances in Experimental Epistemology*. Bloomsbury, pp. 145–174.

Buckwalter, Wesley. (2014b). Intuition fail: Philosophical activity and the limits of expertise. *Philosophy and Phenomenological Research* 92: 378–410.

Buckwalter, Wesley. (2024). A guide to thought experiments in epistemology. In Blake Roeber, Mathius Steup, John Turri, and Ernest Sosa (eds.), *Contemporary Debates in Epistemology, 3rd Edition*. Blackwell, pp. 209–217.

Buckwalter, Wesley, and Jonathan Schaffer. (2015). Knowledge, stakes, and mistakes, *Noûs* 49: 201–234.

Buckwalter, Wesley, and Stephen Stich. (2014). Gender and philosophical intuition. In Shaun Nichols and Joshua Knobe (eds.), *Experimental Philosophy, Volume 2*. Oxford University Press, pp. 307–316.

Byrd, N. (2021). Reflective reasoning & philosophy. *Philosophy Compass* 16: e12786.

Byrd, Nick. (2022). Bounded reflectivism and epistemic identity. *Metaphilosophy* 53: 53–69.

Byrd, Nick. (2023a). Great minds do not think alike: Philosophers' views predicted by reflection, education, personality, and other demographic differences. *Review of Philosophy and Psychology* 14: 647–684.

Byrd, N. (2023b). Reflection-philosophy order effects and correlations: Aggregating and comparing results from mTurk, CloudResearch, Prolific, and undergraduate samples. PsyArXiv.

Byrd, Nick, Brianna Joseph, Gabriela Gongora, and Miroslav Sirota. (2023). Tell us what you really think: A think aloud protocol analysis of the verbal cognitive reflection test. *Journal of Intelligence* 11 . https://doi.org/10.3390/jintelligence11040076

Caie, Michael, Jeremy Goodman, and Harvey Lederman. (2020). Classical opacity. *Philosophy and Phenomenological Research* 101: 524–566.

Cameron, Ross. (2022). *Chains of Being: Infinite Regress, Circularity, and Metaphysical Explanation*. Oxford University Press.

Camp, Elizabeth. (2009). Two varieties of literary imagination: Metaphor, fiction, and thought experiments. *Midwest Studies in Philosophy* 33: 107–130.

Cappelen, Herman. (2012). *Philosophy Without Intuitions*. Oxford University Press.

Cappelen, Herman. (2017). Disagreement in philosophy: An optimistic perspective. In Giuseppina D'Oro and Soren Overgaard (eds.), *Cambridge Companion to Philosophical Methodology*, pp. 56–74.

Carr, Jennifer Rose. (2013). Epistemic utility theory and the aim of belief. *Philosophy and Phenomenological Research* 95: 511–534.

Chalmers, David. (2014). Intuitions in philosophy: A minimal defense. *Philosophical Studies* 171: 535–544.

Chalmers, David. (2015). Why isn't there more progress in philosophy? *Philosophy* 90: 3–31.

Chang, Hasok. (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press.

Chisholm, Roderick. (1989). *Theory of Knowledge*, 3rd Edition. Prentice-Hall.

Church, Alonzo. (1936). An unsolvable problem of elementary number theory. *American Journal of Mathematics* 58: 345–363.

Church, Alonzo. (1940). A formulation of the simple theory of types. *Journal of Symbolic Logic* 5: 56–68.

Cikara, Mina, Rachel Farnsworth, Lasagna Harris, and Susan Fiske. (2010). On the wrong side of the trolley track: Neural correlates of relative social valuation. *Social Cognitive and Affective Neuroscience* 5: 404–413.

Climenhaga, Nevin. (2018). Intuitions are used as evidence in philosophy. *Mind* 127: 69–104.

Cohen, Paul. (1963). The independence of the continuum hypothesis. *Proceedings of the National Academy of Sciences* 50: 1143–1148.

Cohen, Stewart. (1999). Contextualism, skepticism, and the structure of reasons. *Philosophical Perspectives* 13: 587–589.

Cohen, Stewart. (2012). Does practical rationality constrain epistemic rationality? *Philosophy and Phenomenological Research* 85: 447–455.

Cohen, Ted. (1972). The possibility of art: Remarks on a proposal by Dickie. *The Philosophical Review* 82: 69–82.

Colaço, David. (2020). Recharacterizing scientific phenomena. *European Journal for Philosophy of Science* 10: 14.

Colaço, David, Wesley Buckwalter, Stephen Stich, and Edouard Machery. (2014). Epistemic intuitions in fake barn thought experiments. *Episteme* 11: 199–212.

Colaço, David, and Edouard Machery. (2017). The intuitive is a red herring. *Inquiry* 60: 403–419.

Conte, Sebastian. (2022). Are intuitions treated as evidence? Cases from political philosophy. *Journal of Political Philosophy* 30: 411–433.

Costa, Albert, Alice Foucart, Sayuri Hayakawa, Melina Aparici, Jose Apesteguia, Joy Heafner, and Boaz Keysar. Your morals depend on language. *PLOS ONE* 9: e94842.

Cova, Florian. (2016). The folk concept of intentional action: Empirical approaches. In Justin Sytsma and Wesley Buckwalter (eds.), *Blackwell Companion to Experimental Philosophy*, pp. 117–141.

Cova, Florian. (2024). The need for measure calibration in experimental philosophy: The case of measures of folk objectivism. In Joshua Knobe and Shaun Nichols (eds.), *Oxford Studies in Experimental Philosophy, Volume 5*. Oxford University Press, pp. 72–106.

Cova, Florian, Anthony Lantian, and Jordane Boudesseul. (2016). Can the Knobe Effect be explained away? Methodological controversies in the study of the relationship between intentionality and morality. *Personality and Social Psychology Bulletin* 42: 1295–1308.

Craig, William. (1953). On axiomatizability within a system. *Journal of Symbolic Logic* 18: 30–32.
Craig, William. (1957). Three uses of the Herbrand-Gentzen theorem in relating model theory and proof theory. *Journal of Symbolic Logic* 22: 269–285.
Craver, Carl, and Talia Dan-Cohen. (2024). Experimental artifacts. *British Journal for the Philosophy of Science* 75: 253–274.
Cummins, Robert. (1998). Reflections on reflective equilibrium. In Michael DePaul and William Ramsey (eds.), *Rethinking Intuition: The psychology of intuition and its role in philosophical inquiry*. Rowman & Littlefield, pp. 113–128.
de Heide, Rianne, and Peter Grünwald. (2021). Why optional stopping can be a problem for Bayesians. *Psychonomic Bulletin & Review* 28: 795–812.
Daly, Christopher. (2017). *Persistent philosophical disagreement. Proceedings of the Aristotelian Society* 117: 23–40.
Dawes, Robyn, David Faust, and Paul Meehl. (1989). Clinical versus actuarial judgment. *Science* 243: 1668–1674.
Dellsen, Finnur, Insa Lawler, and James Norton. (2021). Thinking about progress: From science to philosophy. *Nous* 56: 814–840.
Demaree-Cotton, Joanna. (2016). Do framing effects make moral intuitions unreliable? *Philosophical Psychology* 29: 1–22.
Dennett, Daniel. (2006). Higher-order truth about chmess. *Topoi* 25: 39–41.
DeRose, Keith. (1992). Contextualism and knowledge attributions. *Philosophy and Phenomenological Research* 52: 913–929.
Deroy, Ophelia. (2014). Modularity of perception. In Mohan Matten (ed.), *The Oxford Handbook of Philosophy of Perception*. Oxford University Press.
Deutsch, Max. (2009). Experimental philosophy and the theory of reference. *Mind & Language* 4: 445–466.
Deutsch, Max. (2010). Intuitions, counter-examples, and experimental philosophy. *Review of Philosophy and Psychology* 1: 447–460.
Deutsch, Max. (2015). *The Myth of the Intuitive*. Oxford University Press.
Deutsch, Max. (2016). Gettier's method. In Jennifer Nado (ed.), *Advances in Experimental Philosophy and Philosophical Methodology*. New York: Bloomsbury, pp. 69–98.
Devitt, Michael. (2006). Intuitions in linguistics. *British Journal for the Philosophy of Science* 57: 481–513.
Devitt, Michael. (2011). Experimental semantics. *Philosophy and Phenomenological Research* 82: 418–435.
Devitt, Michael, and Brian Porter. (2021). Testing the reference of biological kind terms. *Cognitive Science* 45: e12979.
Dietrich, Eric. (2011). There is no progress in philosophy. *Essays in Philosophy* 12: 329–344.
Dinges, Alexander, and Julia Zakkou. (2021). Much at stake in knowledge. *Mind & Language* 31: 729–749.
Donnellan, Keith. (1970). Reference and definite descriptions. *The Philosophical Review* 75: 281–304.
Dorr, Cian. (2016). To be F is to be G. *Philosophical Perspectives* 30: 39–134.
Douglas, Heather. (2000). Inductive risk and values in science. *Philosophy of Science* 67: 559–579.
Douglas, Heather. (2009). *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.
Dunaway, Billy, Anna Edmonds, and David Manley. (2013). The folk probably do think what you think they think. *Australasian Journal of Philosophy* 91: 421–441.
Eliot, T. S. (1920). The perfect critic. In *The Sacred Wood: Essays on Poetry and Criticism*. Methuen & Co., pp. 1–16.
Elliott, Kevin, and Ted Richards (eds.). (2017). *Exploring Inductive Risk: Case Studies of Values in Science*. Oxford University Press.
Elliott, Kevin, and Daniel Steel (eds.). (2017). *Current Controversies in Values and Science*. Routledge.
Engel, Mylan. (2022). Evidence, epistemic luck, reliability, and knowledge. *Acta Analytica* 37: 33–56.

Fantl, Jeremy, and Matthew McGrath. (2002). Evidence, pragmatics, and justification. *The Philosophical Review* 111: 67–94.
Fantl, Jeremy, and Matthew McGrath. (2010). *Knowledge in an Uncertain World*. Oxford University Press.
Feest, Uljana. (2022). Data quality, experimental artifacts, and the reactivity of the psychological subject matter. *European Journal for Philosophy of Science* 12: 1–25.
Feltz, Adam, and Edward Cokeley. (2009). Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition* 18: 342–350.
Feltz, Adam, and Edward Cokeley. (2019). Extraversion and compatibilist intuitions: A ten-year retrospective and meta-analyses. *Philosophical Psychology* 32: 388–403.
Feltz, Adam, and Chris Zarpentine. (2010). Do you know more when it matters less? *Philosophical Psychology* 23: 683–706.
Fischer, Eugen. (2023). Critical ordinary language philosophy: A new project in experimental philosophy. *Synthese* 201: 102.
Fitch, Frederic. (1952). *Symbolic Logic: An Introduction*. The Ronald Press Company.
Fodor, Jerry. (1964). On knowing what we would say. *The Philosophical Review* 198–212.
Fodor, Jerry. (1997). Special sciences: Still autonomous after all these years. *Philosophical Perspectives* 11: 149–163.
Fodor, Jerry. (1998). *Concepts: Where Cognitive Science Went Wrong*. Clarendon Press.
Fox, John. (2008). *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications.
Francis, Kathryn, Philip Beaman, and Nat Hansen. (2019). Stakes, scales, and skepticism. *Ergo: An Open Access Journal of Philosophy* 6: 427–487.
Frankfurt, Harry. (1969). Alternative possibilities and moral responsibility. *The Journal of Philosophy* 66: 829–839.
Frege, Gottloeb. (1893/1952). On sense and reference. In Peter Geach and Max Black (eds.), *Translations from the Philosophical Writings of Gottlob Frege*. Philosophical Library, pp. 56–78.
Friedman, Jane. (2020). The epistemic and the zetetic. *Philosophical Review* 129: 501–536.
Geach, Peter. (1969). The perils of Pauline. *The Review of Metaphysics* 23: 287–300.
Gendler, Tamar, and John Hawthorne. (2005). The real guide to fake barns: A catalogue of gifts for your epistemic enemies. *Philosophical Studies* 124: 331–352.
Gentzen, Gerhard. (1935). Untersuchungen über das logische Schließen. *I Mathematische Zeitschrift* 39 1. doi:10.1007/BF01201353
Gerken, Mikkel. (2013). Epistemic focal bias. *Australasian Journal of Philosophy* 9: 41–61.
Gerken, Mikkel. (2017). *On Folk Epistemology. How We Think and Talk about Knowledge*. Oxford University Press.
Gerken, Mikkel, and James Beebe. (2016). Knowledge in and out of contrast. *Noûs* 50: 133–164.
Gerken, Mikkel, Chad Gonnerman, Joshua Alexander, and John Waterman. (2020). Salient alternatives in perspective. *Australasian Journal of Philosophy* 98: 792–810.
Gettier, Edmund. (1963). Is justified true belief knowledge? *Analysis* 23: 121–123.
Gigerenzer, Gerd, and Henry Brighton. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science* 1: 107–143.
Gödel, Kurt. (1931). Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I. *Monatshefte für Mathematik und Physik* 38: 173–198.
Godfrey-Smith, Peter. (2006). Theories and models in metaphysics. *Harvard Review of Philosophy* 14: 4–19.
Goldin, Claudia, and Cecilia Rouse. (2001). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review* 90: 715–741.
Goldman, Alvin. (1967). A causal theory of knowing. *The Journal of Philosophy* 64: 357–372.
Goldman, Alvin. (1979). What is justified true belief? In George Pappas (ed.), *Justification and Knowledge*. D. Reidel, pp. 1–23.
Goldman, Alvin. (2007). Philosophical intuitions: Their target, their source, and their epistemic status. *Grazer Philosophische Studien* 74: 1–26.

Goldman, Alvin. (2017). Gettier and the epistemic appraisal of philosophical intuition. In Rodrigo Borges, Claudio de Almedia, and Peter Klein (eds.), *Explaining Knowledge: New Essays on the Gettier Problem*. Oxford University Press, pp. 213–230.

Goldman, Alvin, and Joel Pust. (1998). Philosophical theory and intuitional evidence. In Michael DePaul and William Ramsey (eds.), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Rowman & Littlefield, pp. 179–200.

Gonnerman, Chad, Banjit Singh, and Grant Toomey. (2023). Authentic and apparent evidence Gettier cases across American and Indian nationalities. *Review of Philosophy and Psychology* 14: 685–709.

Grice, Paul. (1975). Logic and conversation. In Peter Cole and Jerry Morgan (eds.), *Syntax and Semantics, Vol. 3, Speech Acts*. Academic Press, pp. 41–58.

Grindrod, Jumbly, James Andow, and Nat Hansen. (2019). Third-person knowledge ascriptions: A crucial experiment for contextualism. *Mind & Language* 34: 158–182.

Grundmann, Thomas. (2010). Some hope for intuitions: A reply to Weinberg. In Joachim Horvath and Thomas Grundmann (eds.), *Experimental Philosophy and its Critics*. Routledge, pp. 199–228.

Gutting, Gary. (2016). *Philosophical progress*. In Herman Cappelen, Tamar Gendler, and John Hawthorne (eds.), *The Oxford Handbook of Philosophical Methodology*. Oxford University Press, pp. 309–325.

Haidt, Jonathan. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108: 814–834.

Hales, Steven. (2006). *Relativism and the Foundations of Philosophy*. MIT Press.

Hannon, Michael. (2018). Intuitions, reflective judgments, and experimental philosophy. *Synthese* 195: 4147–4168.

Hansen, Nat. (2018). Review of Avner Baz. The Crisis of Method in Contemporary Analytic Philosophy. *Mind* 128: 936–970.

Hansen, Thorsten, and Karl Gegenfurtner. (2006). Color scaling of discs and natural objects at different luminance levels. *Visual Neuroscience* 23: 603–610.

Hansen, Thorsten, Maria Olkkonen, Sebastian Walter, and Karl Gegenfurtner. (2006). Memory modulates color appearance. *Nature Neuroscience* 9: 1367–1368.

Harding, Sandra. (2015). *Objectivity and Diversity: Another Logic of Scientific Research*. University of Chicago Press.

Haslanger, Sally. (2000). Gender and race: (What) are they? (What) do we want them to be? *Nous* 34: 31–55.

Haukioja, Jussi, Jeske Toorman, Giosuè Baggio, and Jussi Jylkkäc. (2023). Are natural kind terms ambiguous? *Cognitive Science* 47: e13335.

Hawthorne, John. (2003). *Knowledge and Lotteries*. Oxford University Press.

Hawthorne, John, and Jason Stanley. (2008). Knowledge and action. *Journal of Philosophy* 105: 571–590.

Hempel, Carl. (1965). Science and human values. In Carl Hempel, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press, pp. 81–96.

Henderson, David, and Terry Horgan. (2001). The a priori isn't all it is cracked up to be, but it is something. *Philosophical Topics* 29: 219–250.

Henkin, Leon. (1950). Completeness in the theory of types. *Journal of Symbolic Logic* 15: 81–91.

Henne, Paul, Aleksandra Kulesza, Karla Perez, and Augustana Houcek. (2021). Counterfactual thinking and recency effects in causal judgment. *Cognition*. 10.1016/j.cognition.2021.104708. Epub 2021 Apr 2. PMID: 33819848.

Henrich, Joseph, Steven Heine, and Ara Norenzayan. (2010). The weirdest people in the world? *Behavioral and Brain Sciences* 33: 61–83.

Hetherington, Stephen. (2012). The Gettier-illusion: Gettier-partialism and infallibilism. *Synthese* 188: 217–230.

Hilbert, David, and William Ackermann. (1928). *Principles of Theoretical Logic*. American Mathematical Society.

Hitchcock, Christopher, and Joshua Knobe. (2009). Cause and norm. *Journal of Philosophy* 11: 587–612.
Horowitz, Tamara. (1998). Philosophical intuitions and psychological theory. *Ethics* 108: 367–385.
Horvath, Joachim. (2010). How (not) to react to experimental philosophy. *Philosophical Psychology* 23: 447–480.
Horvath, Joachim. (2022). Mischaracterization reconsidered. *Inquiry*: 1–40. https://doi.org/10.1080/0020174X.2021.2019894
Horvath, Joachim, and Alex Wiegmann. (2016). Intuitive expertise and intuitions about knowledge. *Philosophical Studies* 173: 2701–2726.
Horvath, Joachim, and Alex Wiegmann. (2022). Intuitive expertise in moral judgments. *Australasian Journal of Philosophy* 100: 342–359.
Howson, Colin, and Peter Urbach. (1989). *Scientific Reasoning: The Bayesian Approach*. Open Court.
Ichikawa, Jonathan. (2009). Knowing the intuition and knowing the counterfactual. *Philosophical Studies* 145: 435–443.
Ichikawa, Jonathan. (2010). Explaining away intuitions. *Studia Philosophica Estonica* 2: 94–116.
Ichikawa, Jonathan. (2012). Experimentalist pressure against traditional methodology. *Philosophical Psychology* 25: 743–765.
Ichikawa, Jonathan. (2016). Intuitive evidence and experimental philosophy. In Jennifer Nado (ed.), *Advances in Experimental Philosophy and Philosophical Methodology*. Bloomsbury, pp. 155–174.
Ichikawa, Jonathan, and Benjamin Jarvis. (2009). Thought-experiment intuitions and truth in fiction. *Philosophical Studies* 142: 221–246.
Ichikawa, Jonathan, and Benjamin Jarvis. (2013). *The Rules of Thought*. Oxford University Press.
Ichikawa, Jonathan, Ishani Matira, and Brian Weatherson. (2012). In defense of a Kripkean dogma. *Philosophy and Phenomenological Research* 85: 56–68.
Irikefe, Paul. (2020). A fresh look at the expertise reply to the variation problem. *Philosophical Psychology* 33: 840–867.
Jackman, Henry. (2001). Ordinary language, conventionalism, and a priori knowledge. *Dialectica* 55: 315–325.
Jackson, Alex. (2021). Are knowledge ascriptions sensitive to social context? *Synthese* 199: 8579–8610.
Jackson, Frank. (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford University Press.
James, William. (1896). Will to believe. In William James, *The Will to Believe and Other Essays in Popular Philosophy*. Longman's.
Jeffrey, Richard. (1956). Valuation and acceptance of scientific hypotheses. *Philosophy of Science* 22: 237–246.
Jeffrey, Richard. (1965). *The Logic of Decision*. University of Chicago Press.
Jenkins, Carrie. (2008). *Grounding Concepts: An Empirical Basis for Arithmetic Knowledge*. Oxford University Press.
Johnston, Mark, and Sarah-Jane Leslie. (2012). Concepts, analysis, generics, and the Canberra Plan. *Philosophical Perspectives* 26: 113–171.
Kaiser, Marie, and Beate Krickel. (2017). The metaphysics of constructive mechanistic phenomena. *The British Journal for the Philosophy of Science* 68: 745–779.
Kaplan, Mark. (1985). It's not what you know that counts. *Journal of Philosophy* 82: 350–363.
Kauppinen, Antti. (2007). The rise and fall of experimental philosophy. *Philosophical Explorations* 10: 95–118.
Kelly, Thomas. (2003). Epistemic rationality as instrumental rationality: A critique. *Philosophy and Phenomenological Research* 66: 612–640.
Kelp, Christoph. (2021a). Theory of inquiry. *Philosophy and Phenomenological Research* 103: 359–384.
Kelp, Christoph. (2021b). *Inquiry, Knowledge, and Understanding*. Oxford University Press.

Kim, Minsun, and Yuan, Yuan. (2015). No cross-cultural differences in the Gettier car case intuition: A replication study of Weinberg et al. 2001. *Episteme* 12: 355–361.
Kitcher, Philip. (1978). On appealing to the extraordinary. *Metaphilosophy* 9: 99–107.
Kitcher, Philip. (1993). *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford University Press.
Kitcher, Philip. (2001). *Science, Truth, and Democracy*. Oxford University Press.
Kleene, Stephen. (1938). On notation for ordinal numbers. *Journal of Symbolic Logic* 3: 150–155.
Kneer, Marcus, David Colaço, Joshua Alexander, and Edouard Machery. (2022). On second thought: Reflections on the reflection defense. In Joshua Knobe, Tania Lombrozo, and Shaun Nichols (eds.), *Oxford Studies in Experimental Philosophy*, *Volume Four*. Oxford University Press, pp. 257–296.
Knobe, Joshua. (2003). Intentional action and side effects in ordinary language. *Analysis* 63: 190–193.
Knobe, Joshua. (2007). Experimental philosophy and philosophical significance. *Philosophical Explorations* 10: 119–121.
Knobe, Joshua. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences* 33: 315–365.
Knobe, Joshua. (2016). Experimental philosophy is cognitive science. In Justin Sytsma and Wesley Buckwalter (eds.), *A Companion to Experimental Philosophy*. Blackwell, pp. 37–52.
Knobe, Joshua. (2019). Philosophical intuitions are surprisingly robust across demographic differences. *Epistemology and the Philosophy of Science* 56: 29–36.
Knobe, Joshua. (2021). Philosophical intuitions are surprisingly stable across both demographic groups and situations. *Filozofia Nauki* 29: 11–76.
Knobe, Joshua. (2023). Differences and robustness in the patterns of philosophical intuition across demographic groups. *Review of Philosophy and Psychology* 14: 435–455.
Knobe, Joshua, and Shaun Nichols. (2008a). *Experimental Philosophy, Volume One*. Oxford: Oxford University Press.
Knobe, Joshua, and Shaun Nichols. (2008b). An experimental philosophy manifesto. In Joshua Knobe and Shaun Nichols (eds.), *Experimental Philosophy, Volume One*. Oxford University Press, pp. 3–14.
Knobe, Joshua, and Richard Samuels. (2013). Thinking like a scientist: Innateness as a case study. *Cognition* 126: 72–86.
Kornblith, Hilary. (1998). The role of intuitions in philosophical inquiry: An account with no unnatural ingredients. In Michael DePaul and William Ramsey (eds.), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Rowman & Littlefield, pp. 129–142.
Kornblith, Hilary. (2007). Naturalism and Intuitions. *Grazer Philosophische Studien* 72: 27–49.
Kornblith, Hilary. (2010). What reflective endorsement cannot do. *Philosophy and Phenomenological Research* 80: 1–19.
Kripke, Saul. (1963). Semantical considerations on modal logic. *Acta Philosophica Fennica* 16: 83–94.
Kripke, Saul. (1980). *Naming and Necessity*. Harvard University Press.
Kumar, Victor, and Joshua May. (2019). How to debunk moral beliefs. In Jussi Suikkanen and Antti Kauppinen (eds.), *Methodology and Moral Philosophy*. Routledge, pp. 25–48.
Kutner, Michael, Christopher Nachtsheim, John Neter, and William Li. (2004). *Applied Linear Statistical Models*. McGraw-Hill.
Lam, Barry. (2010). Are Cantonese speakers really descriptivists? Revisiting cross-cultural semantics. *Cognition* 115: 320–329.
Landes, Ethan. (2020). The threat of the intuition-shaped hole. *Inquiry* 66(4): 539–564.
Landes, Ethan. (2023). Philosophical producers, philosophical consumers, and the metaphilosophical value of original texts. *Philosophical Studies* 180: 207–225.
Lehrer, Keith. (1990). *Theory of Knowledge*. Routledge Press.
Leonelli, Sabina. (2009). On the locality of data and claims about phenomena. *Philosophy of Science* 76: 737–749.
Levi, Isaac. (1962). On the seriousness of mistakes. *Philosophy of Science* 29: 47–65.

Levin, Janet. (2005). The evidential status of philosophical intuition. *Philosophical Studies* 121: 193–224.
Levin, Janet. (2019). A case for the method of cases. *Philosophy and Phenomenological Research* 98: 230–238.
Levy, Neil. (2021). *Bad Beliefs: Why They Happen to Good People*. Oxford University Press.
Lewis, David. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy* 50: 249–258.
Lewis, David. (1983). *Philosophical Papers, Volume 1*. Oxford University Press.
Lewis, David. (1996). Elusive knowledge. *Australasian Journal of Philosophy* 74: 549–567.
Liao, Shen-yi. (2016). Are philosophers good intuition predictors? *Philosophical Psychology* 29: 1004–1014.
Liao, S. Matthew. (2008). A Defense of Intuitions. *Philosophical Studies* 140: 247–262.
Liao, S. Matthew, Alex Wiegmann, Joshua Alexander, and Gerard Vong. (2012). Putting the trolley in order: Experimental philosophy and the loop case. *Philosophical Psychology* 25: 661–671.
Loftus, Elizabeth. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of the mind. *Learning & Memory* 12: 361–366.
Löhr, Guido. (2019). The experience machine and the expertise defense. *Philosophical Psychology* 32: 257–273.
Lombrozo, Tania. (2013). The ironic success of experimental philosophy. https://www.npr.org/sections/13.7/2013/03/23/175145568/the-ironic-success-of-experimental-philosophy
Longino, Helen. (1990). *Science as Social Knowledge*. Princeton University Press.
Lopes, Dominic. (2008). Nobody needs a theory of art. *Journal of Philosophy* 105: 109–127.
Ludwig, Kirk. (2007). The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy* 31: 128–159.
Machery, Edouard. (2012). Expertise and intuitions about reference. *Theoria* 27: 37–54.
Machery, Edouard. (2017). *Philosophy Within Its Proper Bounds*. Oxford University Press.
Machery, Edouard. (2020). Cross-Cultural Semantics at 15. In Stephen Biggs and Heimer Geirsson (eds.), *The Routledge Handbook of Linguistic Reference*. Routledge, pp. 535–550.
Machery, Edouard. (2023). Why Variation Matters to Philosophy. *Res Philosophica* 100: 1–22.
Machery, Edouard and Stephen Stich. (2013). The role of experiment. In Gillian Russell and Delia Graff Fara (eds.), *Routledge Companion to Philosophy of Language*. Routledge, pp. 513–530.
Machery, Edouard, Max Deutsch, Justin Sytsma, Ron Mallon, Shaun Nichols, and Stephen Stich. (2010). Semantic intuitions: Reply to Lam. *Cognition* 117: 361–366.
Machery, Edouard, Ron Mallon, Shaun Nichols, and Stephen Stich. (2004). Semantics, cross-cultural style. *Cognition* 92: B1–B12.
Machery, Edouard, Christopher Olivola, and Molly De Blanc. (2009). Linguistic and metalinguistic intuitions in the philosophy of language. *Analysis* 69: 689–694.
Machery, Edouard, Stephen Stich, David Ros, Amita Chatterjee, Kaori Karasawa, Noel Struchiner, Smita Sirker, Naoki Usui, and Takaaki Hashimoto. (2017). Gettier across cultures. *Noûs* 51: 645–664.
Machery, Edouard, Stephen Stich, David Ros, Amita Chatterjee, Kaori Karasawa, Noel Struchiner, Smita Sirker, Naoki Usui, and Takaaki Hashimoto. (2018). Gettier was framed! In Masaharu Mizumoto, Stephen Stich, and Eric McCready (eds.), *Epistemology for the Rest of the World*. Oxford University Press, pp. 123–148.
Machery, Edouard, Justin Sytsma, and Max Deutsch. (2015). Speaker's reference and cross-cultural semantics. In Andrea Bianchi (ed.), *On Reference*. Oxford University Press, pp. 62–76.
Macpherson, Fiona. (2012). Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research* 84: 24–62.
Malle, Bertram, and Joshua Knobe. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology* 33: 101–121.
Mallon, Ron. (2017). Intuitive diversity and disagreement. In Jennifer Nado (ed.), *Advances in Experimental Philosophy and Philosophical Methodology*. Bloomsbury, pp. 99–124.
Mallon, Ron, Edouard Machery, Shaun Nichols, and Stephen Stich. (2009). Against arguments from reference. *Philosophy and Phenomenological Research* 79: 332–356.

Malmgren, Anna-Sara. (2011). Rationalism and the content of intuitive judgments. *Mind* 120: 263–327.
Martí, Genoveva. (2009). Against semantic multi-culturalism. *Analysis* 69: 42–8.
Martí, Genoveva. (2020). Experimental semantics, descriptivism and anti-descriptivism. Should we endorse referential pluralism? In Andrea Bianchi (ed.), *Language and Reality from a Naturalistic Perspective. Philosophical Studies Series* 142. Springer.
Mates, Benson. (1958). On the verification of statements about ordinary language. *Inquiry* 1: 1–4.
Mates, Benson. (1964). *Elementary Logic*. Oxford University Press.
Matheson, Jonathan. (2024). Why think for yourself? *Episteme* 21: 320–338.
May, Joshua, Walter Sinnott-Armstrong, Jay G. Hull, and Aaron Zimmerman. (2010). Practical interests, relevant alternatives, and knowledge attributions: An empirical study. *Review of Philosophy and Psychology* 1: 265–273.
Maynes, Jeffrey. (2021). The method(s) of cases. *Philosophical Psychology* 34: 102–124.
McKenna, Michael. (2014). Resisting the manipulation argument: A hard-liner takes it on the chin. *Philosophy and Phenomenological Research* 89: 467–484.
Mercier, Hugo. (2011). On the universality of argumentative reasoning. *Journal of Cognition and Culture* 11: 85–113.
Mercier, Hugo. (2017). Confirmation bias—myside bias. In Rüdiger F Pohl (ed.), *Cognitive Illusions: Intriguing Phenomena in Thinking, Judgment, and Memory*. Routledge, pp. 99–114.
Mercier, Hugo, and Daniel Sperber. (2017). *The Enigma of Reason*. Harvard University Press.
Michaelian, Kourken, Dorothea Debus, and Denis Perrin. (2018). *New Directions in the Philosophy of Memory*. Routledge.
Murray, Dylan, and Eddy Nahmias. (2014). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research* 88: 434–467.
Nadelhoffer, Thomas. (2004). Blame, badness, and intentional action: A reply to Knobe and Mendlow. *The Journal of Theoretical and Philosophical Psychology* 24: 259–269.
Nadelhoffer, Thomas. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. Philosophical Explorations 9: 203–219.
Nadelhoffer, Thomas, and Adam Feltz. (2008). The actor-observer bias and moral intuitions: Adding fuel to Sinnott-Armstrong's fire. *Neuroethics* 1: 133–144.
Nado, Jennifer. (2014). Why intuition? *Philosophy and Phenomenological Research* 89: 15–41.
Nado, J. (2015). Intuition, philosophical theorizing, and the threat of skepticism. In Eugene Fischer and John Collins (eds.), *Experimental Philosophy, Rationalism, and Naturalism: Rethinking Philosophical Method*. Routledge, pp. 204–221.
Nado, Jennifer. (2016a). Experimental philosophy 2.0. *Thought* 5: 159–168.
Nado, Jennifer. (2016b). The intuition deniers. *Philosophical Studies* 173: 781–800.
Nado, Jennifer. (2017). Knowledge second (for metaphilosophy). In A. Coliva and N. Pedersen (eds.), *Epistemic Pluralism*. Palgrave, pp. 145–170.
Nagel, Jennifer. (2010). Knowledge ascriptions and the psychological consequences of thinking about error. *The Philosophical Quarterly* 60: 286–306.
Nagel, Jennifer. (2012). Intuitions and experiments: A defense of the case method in epistemology. *Philosophy and Phenomenological Research* 85: 495–527.
Nagel, Jennifer, Valarie San Juan, and Raymond Mar. (2013). Lay denial of knowledge for justified true beliefs. *Cognition* 129: 652–661.
Nahmias, Eddy, Justin Coates, and Trevor Kvaran. (2007). Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest Studies in Philosophy* 31: 214–242.
Nahmias, Eddy, and Dylan Murray. (2011). Experimental philosophy on free will: An error theory for incompatibilist intuitions. In Jesus Aguilar, Andrei Buckareff, and Keith Frankish (eds.), *New Waves in Philosophy of Action*. Palgrave Macmillan, pp. 189–216.
Nahmias, Eddy, Stephen Morris, Thomas Nadelhoffer, and Jason Turner. (2006). Is incompatibilism intuitive? *Philosophy and Phenomenological Research* 73: 28–53.
Nichols, Shaun, and Michael Bruno. (2010). Intuitions about personal identity: An empirical study. *Philosophical Psychology* 23: 293–312.

Nichols, Shaun, and Joshua Knobe. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous* 4: 663–685.
Nichols, Shaun, N., Angel Pinillos, and Ron Mallon. (2016). Ambiguous reference. *Mind* 125: 145–175.
Nisbett, Richard. (2003). *The Geography of Thought*. Free Press.
Nisbett, Richard, Kaiping Peng, Incheol Choi, and Ara Zorensayan. (2001). Culture and systems of thought: Holistic vs. analytic cognition. *Psychological Review* 108: 291–310.
Nisbett, Richard, and Timothy Wilson. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology* 35: 250–256.
O'Callaghan, Casey. (2011). Perception and multimodality. In Eric Margolis, Richard Samuels, and Stephen Stich (eds.), *Oxford Handbook of Philosophy of Cognitive Science*. Oxford University Press, pp. 92–117.
Osborne, Philip. (2014). A modest response to empirical skepticism about intuitions. *Episteme* 11: 443–456.
Papineau, David. (2009). The poverty of conceptual analysis. *Aristotelian Society Supplementary Volume* 83: 1–30.
Parfit, Derek. (1984). *Reasons and Persons*. Clarendon Press.
Parsons, Charles. (1995). Platonism and mathematical intuition in Kurt Gödel's thought. *Bulletin of Symbolic Logic* 1: 44–74.
Pastötter, Barnhard, Sabine Gleixner, Theresa Neuhauser, and Karl-Heinz T. Bäuml. (2013). To push or not to push? Affective influences on moral judgment depend on decision frame. *Cognition* 126: 373–377.
Paul, L. A. (2012). Metaphysics as modeling: The handmaiden's tale. *Philosophical Studies* 160: 1–29.
Perry, John. (1972). Can the self divide? *The Journal of Philosophy* 69: 463–488.
Petrinovich, Lewis, and Patricia O'Neill. (1996). Influence of wording and framing effects on moral intuitions. *Ethology & Sociobiology* 17: 145–171.
Pettigrew, Richard. (2011). An improper introduction to epistemic utility theory. In Henk de Regt, Samir Okasha, and Stephen Hartmann (eds.), *Proceedings of EPSA: Amsterdam '09*. Springer, pp. 287–301.
Pettit, Derek, and Joshua Knobe. (2009). The pervasive impact of moral judgment. *Mind & Language* 24: 586–604.
Pinillos, Ángel. (2012). Knowledge, experiments, and practical interests. In Jessica Brown and Mikkel Gerken (eds.), *Knowledge Ascriptions*. Oxford University Press, pp. 192–219.
Plantinga, Alvin. (1993). *Warrant and Proper Function*. Oxford University Press.
Porter, Brian, Kelli Barr, Abdellatif Bencherifa, Wesley Buckwalter, Yasuo Deguchi, Emanuele Fabiano, Takaaki Hashimoto, Julia Halamova, Joshua Homan, Kaori Karasawa, Martin Kanovsky, Hackjin Kim, Jordan Kiper, Minha Lee, Xiaofei Liu, Veli Mitova, Rukmini Bhaya, Ljiljana Pantovic, Pablo Quintanilla, Josien Reijer, Pedro Romero, Purmina Singh, Salma Tber, Daniel Wilkenfeld, Stephen Stich, Clark Barrett, and Edouard Machery. (2024). A puzzle about knowledge ascriptions. *Nous*. https://doi.org/10.1111/nous.12515.
Powell, Derek, Zachary Horne, Ángel Pinillos, Keith Holyoak. (2013). Justified true belief triggers false recall of "knowing." *Proceedings of the Annual Meeting of the Cognitive Science Society* 35. https://escholarship.org/uc/item/1dq4g72z
Putnam, Hilary. (1975). Explanation and reference. In *Mind, Language, and Reality*. Cambridge University Press, pp. 196–214.
Putnam, Hilary. (1990). *Realism with a Human Face*. Cambridge, MA: Harvard University Press.
Ramsey, William. (2019). Intuitions as evidence facilitators. *Metaphilosophy* 50: 76–99.
Rose, David, and Shaun Nichols. (2013). The lessons of bypassing. *Review of Philosophy and Psychology* 4: 599–619.
Rose, David, Edouard Machery, Stephen Stich, Mario Alai, Adriano Angelucci, Renatas Berniūnas, Emma E. Buchtel, Amita Chatterjee, Hyundeuk Cheon, In-Rae Cho, Daniel Cohnitz, Florian Cova, Vilius Dranseika, Ángeles Eraña Lagos, Laleh Ghadakpour, Maurice Grinberg, Ivar Hannikainen, Takaaki Hashimoto, Amir Horowitz, Evgeniya Hristova,

Yasmina Jraissati, Veselina Kadreva, Kaori Karasawa, Hackjin Kim, Yeonjeong Kim, Minwoo Lee, Carlos Mauro, Masaharu Mizumoto, Sebastiano Moruzzi, Christopher Y. Olivola, Jorge Ornelas, Barbara Osimani, Carlos Romero, Alejandro Rosas Lopez, Massimo Sangoi, Andrea Sereni, Sarah Songhorian, Paulo Sousa, Noel Struchiner, Vera Tripodi, Naoki Usui, Alejandro Vázquez del Mercado, Giorgio Volpe, Hrag Abraham Vosgerichian, Xueyi Zhang, and Jing Zhu. (2019). Nothing at stake in knowledge. *Nous* 53: 224–247.

Rouder, Jeffrey. (2014). Optional stopping: No problem of Bayesians. *Psychonomic Bulletin & Review* 21: 301–308.

Roxborough, Craig, and Jill Cumby. (2009). Folk psychological concepts: Causation. *Philosophical Psychology* 22: 205–213.

Rudner, Richard. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science* 20: 1–6.

Russell, Bertrand. (1919). Knowledge by acquaintance and knowledge by description. *Proceedings of the Aristotelian Society* 11: 108–128.

Samuels, Richard, Stephen Stich, and Michael Bishop. (2002). Ending the rationality wars: How to make disputes about human rationality disappear. In Renee Elio (ed.), *Common Sense, Reasoning and Rationality*. Oxford University Press, pp. 236–268.

Sanborn, Adam, and Thomas Hills. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review* 21: 283–300.

Sartwell, Crispin. (1992). Why knowledge is merely true belief. *Journal of Philosophy* 89: 167–180.

Saul, Jennifer. (2007). *Simple Sentences, Substitution, and Intuition*. Oxford University Press.

Schaffer, Jonathan, and Joshua Knobe. (2012). Contrastive knowledge surveyed. *Noûs* 46: 675–708.

Schickore, Jutta. (2019). The structure and function of experimental controls in the life sciences. *Philosophy of Science* 86: 203–218.

Schindler, Samuel, and Pierre Saint-Germier. (2023). Putting philosophical expertise to the test. *Australasian Journal of Philosophy* 101: 592–608.

Schwitzgebel, Eric, and Fiery Cushman. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language* 27: 135–153.

Schwitzgebel, Eric, and Fiery Cushman. (2015). Philosophers' biased judgments persist despite training, expertise, and reflection. *Cognition* 141: 127–137.

Schwitzgebel, Eric, Liam Kofi Bright, Carolyn Dicey Jennings, Morgan Thompson, and Eric Winsberg. (2021). The diversity of philosophy students and faculty. *The Philosophers' Magazine* 93: 71–90.

Searle, John. (1958). Proper Names. *Mind* 67: 166–173.

Segal, Gabriel. (1997). Synaesthesia: Implications for modularity of mind. In John Harrison and Simon Baron-Cohen (eds.), *Synaesthesia: Classic and Contemporary Readings*. Blackwell, pp. 211–223.

Seyedsayamdost, Hamid. (2015). On gender and philosophical intuition: Failure of replication and other negative results. *Philosophical Psychology* 28(5): 642–673.

Seyedsayamdost, Hamid. (2019). Philosophical expertise and philosophical methodology. *Metaphilosophy* 50: 110–129.

Shoemaker, Sydney. (1963). *Self-Knowledge and Self-Identity*. Cornell University Press.

Shope, Robert. (2002). Conditions and analyses of knowing. In Kurt Moser (ed.), *The Oxford Handbook of Epistemology*. Oxford University Press, pp. 25–70.

Sider, Theodore. (2001). Criteria of personal identity and the limits of conceptual analysis. *Philosophical Perspectives* 15: 189–209.

Sider, Theodore. (2010). *Logic for Philosophy*. Oxford University Press.

Siegel, S. (2011). Cognitive penetrability and perceptual justification. *Noûs* 46: 201–222.

Simons, Daniel, and Michael Ambinder. (2005). Change blindness: Theory and consequences. *Current Directions in Psychological Science* 14: 44–48.

Skyrms, Brian. (1966). *Choice & Chance*. Cengage Learning.

Sosa, Ernest. (1998). Minimal intuition. In Michael DePaul and William Ramsey (eds.), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Rowman & Littlefield, pp. 257–270.

Sosa, Ernest. (1999). How to defeat opposition to Moore. *Philosophical Perspectives* 13: 141–153.

Sosa, Ernest. (2007a). Intuitions: Their nature and epistemic efficacy. *Grazer Philosophische Studien* 74: 51–67.

Sosa, Ernest. (2007b). Experimental philosophy and philosophical intuition. *Philosophical Studies* 132: 99–107.

Sosa, Ernest. (2009). A defense of the use of intuitions in philosophy. In Michael Bishop and Daniel Murphy (eds.), *Stich and his Critics*. Wiley-Blackwell, pp. 101–112.

Sripada, Chandra Sekhar, and Jason Stanley. (2012). Empirical tests of interest-relative invariantism. *Episteme* 9: 3–26.

Stanley, Jason. (2005). *Knowledge and Practical Interests*. Oxford University Press.

Stanovich, Keith. (2021). *The Bias That Divides Us: The Science and Politics of Myside Thinking*. MIT Press.

Stanovich, Keith, and Richard West. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology* 94: 672–695.

Stanovich, Keith, Richard West, and Maggie Toplak. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science* 22: 259–264.

Starmans, Christina, and Ori Friedman. (2012). The folk concept of knowledge. *Cognition* 124: 272–283.

Starmans, Christina, and Ori Friedman. (2020). Expert or esoteric? Philosophers attribute knowledge differently than all other academics. *Cognitive Science* 44. 10.1111/cogs.12850

Sterken, Rachael. (2016). Generics, covert structure, and logical form. *Mind & Language* 31(5): 503–529.

Stich, Stephen. (1990). *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation*. MIT Press.

Stich, Stephen. (1998). Reflective equilibrium, analytic epistemology, and the problem of cognitive diversity. In Michael DePaul and William Ramsey (eds.), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Rowman & Littlefield, pp. 95–112.

Stich, Stephen, and Edouard Machery. (2022). Demographic differences in philosophical intuition: A reply to Joshua Knobe. *Review of Philosophy and Psychology* 14: 401–434.

Stoljar, Daniel. (2017). *Philosophical Progress: In Defense of a Reasonable Optimism*. Oxford: Oxford University Press.

Strevens, Michael. (2003). The role of the priority rule in science. *Journal of Philosophy* 100: 55–79.

Strevens, Michael. (2019). Philosophy unbounded. *Philosophy and Phenomenological Research* 98: 239–245.

Swain, Stacey, Joshua Alexander, and Jonathan Weinberg. (2008). The instability of philosophical intuitions: Running hot and cold on Truetemp. *Philosophy and Phenomenological Research* 76: 138–155.

Swinburne, Richard. (1984). Personal identity: The dualist theory. In Sydney Shoemaker and Richard Swinburne (eds.), *Personal Identity: Great Debates in Philosophy*. Blackwell, pp. 1–66.

Sytsma, Justin, and Jonathan Livengood. (2011). A new perspective concerning experiments on semantic intuitions. *Australasian Journal of Philosophy* 89: 315–372.

Sytsma, Justin, and Jonathan Livengood. (2016). *Theory and Practice of Experimental Philosophy*. Broadview Press.

Sytsma, Justin, Jonathan Livengood, Ryoji Sato, and Mineki Oguchi. (2015). Reference in the land of the rising sun: A cross-cultural study on the reference of proper names. *Review of Philosophy and Psychology* 6: 213–230.

Talbot, Brian. (2013). Reforming intuition pumps: When are the old ways the best? *Philosophical Studies* 165: 315–334.

Tarski, Alfred. (1933). "Pojecie prawdy w jezykach nauk dedukeycyjnych," translated as "On the Concept of Truth in Formalized Languages." In Alfred Tarski. (1983). *Logic, Semantics, Metamathematics*, 2nd Edition. Hackett Publishing, pp. 152–278.

Tarski, Alfred. (1936). "On the Concept of Logical Consequence." In Alfred Tarski. (1983). *Logic, Semantics, Metamathematics*, 2nd Edition. Hackett Publishing, pp. 409–420.

Teller, Paul. (2010). "Saving the phenomena" today. *Philosophy of Science* 77: 815–826.

Thompson, Judith Jarvis. (1971). A defense of abortion. *Philosophy & Public Affairs* 1: 47–66.

Thomson, Judith Jarvis. (1985). The trolley problem. *The Yale Law Journal* 94: 1395–1415.

Thorstad, David. (Forthcoming). Norms of inquiry. *Philosophical Topics*.

Titlebaum, Michael. (2022). *Fundamentals of Bayesian Epistemology*. Oxford University Press.

Turing, Alan. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* 42: 230–265.

Turner, Jason. (2016). Curbing enthusiasm about grounding. *Metaphysics* 30: 366–396.

Turri, John. (2013). A conspicuous art: Putting Gettier to the test. *Philosophers' Imprint* 13: 1–16.

Turri, John. (2015). Skeptical appeal: The source-content bias. *Cognitive Science* 39: 307–324.

Turri, John. (2016). How to do better: Toward normalizing experimentation in epistemology. In Jennifer Nado (ed.), *Advances in Experimental Philosophy and Philosophical Methodology*. Bloomsbury, pp. 35–51.

Turri, John. (2017). Epistemic contextualism: An idle hypothesis. *Australasian Journal of Philosophy* 95: 141–156.

Turri, John, Wesley Buckwalter, and David Rose. (2016). Actionability judgments cause knowledge judgments. *Thought: A Journal of Philosophy* 5: 212–222.

Tversky, Amos, and Daniel Kahneman. (1974). Judgment under uncertainty: Heuristics and biases. *Science* 185: 1124–1131.

Tversky, Amos, and Daniel Kahneman. (1981). The framing of decisions and the psychology of choice. *Science* 211: 453–458.

Uhlmann, Eric, David Pizzaro, David Tannenbaum, and Peter Ditto. (2009). The motivated use of moral principles. *Judgment and Decision Making* 4: 476–491.

Unger, Peter. (1990). *Identity, Consciousness, and Value*. Oxford University Press.

Valdesolo, Piercarlo, and David DeSteno. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science* 17: 476–477.

van Dongen, Noah, Matteo Colombo, Felipe Romero, and Jan Sprenger. (2021). Intuitions About the Reference of Propoer Names: A Meta-Analysis. *Review of Philosophy and Psychology* 12: 745–774.

van Inwagen, Peter. (1980). Philosophers and the words "human body". In Peter van Inwagen (ed.), *Time and Causes*. Dordrecht, pp. 283–289.

van Leeuwen, Neil. (2013). The meaning of "imagine" part I: Constructive imagination. *Philosophy Compass* 3: 220–230.

Waterman, J., C. Gonnerman, K. Yan, and J. Alexander. (2018). Knowledge, certainty, and skepticism: A cross-cultural study. In Mazuharo Mizumoto, Stephen Stich, and Eric McCreay (eds.), *Epistemology for the Rest of the World*. Oxford University Press, pp. 187–214.

Weatherson, Brian. (2003). What good are counterexamples? *Philosophical Studies* 115: 1–31.

Weatherson, Brian. (2014). Centrality and marginalization. *Philosophical Studies* 171: 517–533.

Weinberg, Jonathan. (2007). How to challenge intuitions empirically without risking skepticism. *Midwest Studies in Philosophy* 31: 318–343.

Weinberg, Jonathan. (2008). Configuring the cognitive imagination. *New Waves in Aesthetics*: 203–223.

Weinberg, Jonathan. (2009). On doing better, experimental style. *Philosophical Studies* 145: 455–464.

Weinberg, Jonathan. (2014a). Cappelen between rock and a hard place. *Philosophical Studies* 171: 545–553.

Weinberg, Jonathan. (2014b.) The promise of experimental philosophy and the inference to signal. In James Beebe (ed.), *Advances in Experimental Epistemology*. Bloomsbury Academic, pp. 193–207.

Weinberg, Jonathan. (2015). Humans as instruments: Or, the inevitability of experimental philosophy. In Eugen Fischer (ed.), *Method, Rationalism, and Naturalism: The Impact of Experimental Philosophy*. Oxford University Press, pp. 171–187.

Weinberg, Jonathan. (2016a). Intuitions. In Herman Cappelen, Tamar Gendler, and John Hawthorne (eds.), *The Oxford Handbook of Philosophical Methodology*. Oxford: Oxford University Press, pp. 287–308.

Weinberg, Jonathan. (2016b). Experimental philosophy, noisy intuition, and messy inferences. In Jennifer Nado (ed.), *Advances in Experimental Philosophy and Philosophical Methodology*. Bloomsbury, pp. 11–34.

Weinberg, Jonathan. (2016c). Going positive by going negative: Keeping X-phi relevant and dangerous. In Justin Sytsma and Wesley Buckwalter (eds.), *The Blackwell Companion to Experimental Philosophy*. Blackwell, pp. 71–86.

Weinberg, Jonathan. (2018). Are aestheticians' intuitions sitting pretty? In Sebastien Rehault and Florian Cova (eds.), *Advances in Experimental Philosophy of Aesthetics*. Bloomsbury, pp. 267–288.

Weinberg, Jonathan, and Stephen Crowley. (2009). Loose constitutivity and armchair philosophy. *Studia Philosophica Estonica* 2(2): 177–195.

Weinberg, Jonathan, Joshua Alexander, Chad Gonnerman, and Shane Reuter. (2012). Restrictionism and reflection: Challenged deflected, or simply redirected? *The Monist* 95: 200–222.

Weinberg, Jonathan, and Joshua Alexander. (2014). The challenge of sticking with intuitions through thick and thin. In Anthony Booth and Darrell Rowbottom (eds.), *Intuitions*. Oxford University Press, pp. 187–212.

Weinberg, Jonathan, and Steve Crowley. (2009). The X-phi(les): Unusual insights into the nature of inquiry. *Studies in the History and Philosophy of Science, Part A* 40: 227–232.

Weinberg, Jonathan, Stephen Crowley, Chad Gonnerman, Ian Vandewalker, and Stacey Swain. (2012). Intuition and calibration. *Essays in Philosophy* 13(1): 257–284.

Weinberg, Jonathan, Chad Gonnerman, Cameron Buckner, and Joshua Alexander. (2010). Are philosophers expert intuiters? *Philosophical Psychology* 23: 331–355.

Weinberg, Jonathan, Shaun Nichols, and Stephen Stich. (2001). Normativity and epistemic intuitions. *Philosophical Topics* 29: 429–460.

Weinberg, Jonathan, and Ellie Wang. (2010). Naturalism's perils, naturalism's promises: A comment on Appiah's Experiments in Ethics. *Neuroethics* 3: 215–222.

Wiegmann, Alex, Yasmina Okan, and Jonas Nagel. (2012). Order effects in moral judgment. *Philosophical Psychology* 25: 813–836.

Wiegmann, Alex, and Michael Waldmann. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition* 131: 28–43.

Wiegmann, Alex, Joachim Horvath, and Karina Meyer. (2020). Intuitive expertise and irrelevant options. *Oxford Studies in Experimental Philosophy* 3: 275–310.

Williams, Bernard. (1970). The self and the future. *The Philosophical Review* 79(2): 161–180.

Williamson, Timothy. (2004). Philosophical intuitions and scepticism about judgment. *Dialectica* 58: 109–153.

Williamson, Timothy. (2005). Contextualism, subject-sensitive invariantism and knowledge of knowledge. *The Philosophical Quarterly* 55(219): 213–235.

Williamson, Timothy. (2006). Must do better. In Patrick Greenough and Michael Lynch (eds.), *Truth and Realism*. Oxford: Oxford University Press, pp. 278–292.

Williamson, Timothy. (2007). *The Philosophy of Philosophy*. Blackwell Publishing.

Williamson, Timothy. (2009). Precis of The Philosophy of Philosophy. *Philosophical Studies* 145: 431–434.

Williamson, Timothy. (2010). Philosophy vs. imitation philosophy. *The New York Times*.

Williamson, Timothy. (2016). Philosophical criticisms of experimental philosophy. In Justin Sytsma and Wesley Buckwalter (eds.), *Routledge Companion to Experimental Philosophy*. Routledge, pp. 22–36.

Williamson, Timothy. (2017). Model-building in philosophy. In Russell Blackford and Damien Broderick (eds.), *Philosophy's Future: The Problem of Philosophical Progress*. Wiley, pp. 151–179.

Wilson, Jessica. (2014). No work for a theory of grounding. *Inquiry: An Interdisciplinary Journal of Philosophy* 57: 535–579.

Wilson, Timothy. (2004). *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, MA: Harvard University Press.

Woodward, James. (1989). Data and phenomena. *Synthese* 79: 393–472.

Woodward, James. (2000). Data, phenomena, and reliability. *Philosophy of Science* 67: S163–S179.

Woodward, James. (2011). Data and phenomena: A restatement and defense. *Synthese* 182: 165–179.

Wright, Jennifer Cole. (2010). On intuitional stability: The clear, the strong, and the paradigmatic. *Cognition* 115: 491–503.

Yablo, Stephen. (1993). Is conceivability a guide to possibility? *Philosophy and Phenomenological Research* 1–42.

Yablo, Stephen. (2006). No fool's gold: Notes on illusions of possibility. In Manuel Garcia-Carpintero and Josep Macia (eds.), *Two-Dimensional Semantics*. Oxford: Oxford University Press, pp. 327–346.

Zangwill, Nick. (2002). Are the counterexamples to aesthetic theories of art? *Journal of Aesthetics and Art Criticism* 609: 111–116.

Ziółkowski, Adrian. (2021). The stability of philosophical intuitions: Failed replications of Swain et al. *Episteme* 18: 328–346.

Ziółkowski, Adrian, Alex Wiegmann, Joachim Horvath, and Edouard Machery. (2023). Truetemp cooled down: The stability of truetemp intuitions. *Synthese* 201: 1–19.

Index

For the benefit of digital users, indexed terms that span two pages (e.g., 52–53) may, on occasion, appear on only one of those pages.